Analítica de datos y herramientas de inteligencia artificial II

Profesor:

Alfredo García Suárez

Actividad 4. (Extracción de Características)

José Jaime Ponce de León Cobos A01552256

30 de septiembre del 2023

Resumen

En el presente reporte se presenta un análisis de datos llevado a cabo a partir de un archivo que recopila información de diversas tiendas. Estos datos fueron obtenidos a través de una encuesta realizada a varios empleados de tiendas, con el objetivo de comprender mejor su funcionamiento y rendimiento.

El análisis de datos se realizó utilizando programación, lo que permitió una exploración exhaustiva de la información disponible. Uno de los pasos fundamentales en este proceso fue la identificación y tratamiento de datos nulos. La eliminación de datos faltantes y su sustitución con técnicas como el llenado con la media o la interpolación hacia atrás fue esencial para garantizar la integridad de los datos y evitar sesgos en los resultados.

Además, se llevó a cabo una detección y eliminación de valores atípicos, utilizando métodos basados en la desviación estándar y cuantiles. Esto fue crucial para asegurar que los datos utilizados en el análisis fueran representativos y no estuvieran distorsionados por valores atípicos.

En el proceso de análisis, se aplican filtros a las variables para seleccionar las más relevantes y significativas para el estudio. Esto ayudó a enfocar la investigación en aspectos específicos y a reducir la complejidad de los datos.

Finalmente, para comprender mejor visualmente los patrones y tendencias en los datos, se utilizaron gráficos y visualizaciones. Estos gráficos facilitan la identificación de relaciones entre variables, la evolución temporal de los datos y la identificación de posibles áreas de mejora en las tiendas.

Objetivos

1. Aprender y aplicar técnicas de eliminación de datos nulos: El objetivo central del informe es comprender y utilizar técnicas efectivas para abordar los datos faltantes en un conjunto de datos, como el llenado con la media o la interpolación backforward, con el fin de mantener la integridad de los datos y garantizar un análisis preciso.

2. Identificar y eliminar outliers: Otro objetivo importante es aprender a detectar y eliminar valores atípicos o outliers en los datos, utilizando métodos basados en la desviación estándar o cuantiles. Esto es fundamental para obtener resultados más precisos y significativos en el análisis.

3. Aplicar filtros a las variables: Se busca comprender cómo aplicar filtros a las variables para seleccionar las más relevantes y significativas para el análisis. Esto ayuda a simplificar el proceso y enfocarse en aspectos específicos de interés.

4. Interpretar gráficos y visualizaciones: Un objetivo adicional es aprender a utilizar gráficos y visualizaciones para interpretar los datos de manera efectiva. Estos gráficos ayudan a identificar patrones, tendencias y relaciones entre las variables, lo que facilita la comprensión y la toma de decisiones informadas.

Introducción

En el presente informe, se presentarán los resultados de un análisis exhaustivo de datos basado en una base de datos que recopila información de diversas tiendas. A lo largo de este trabajo, se ha llevado a cabo un minucioso proceso de exploración y manipulación de los datos, con el objetivo de comprender mejor el funcionamiento y rendimiento de estas tiendas.

Una parte esencial de este análisis implica la visualización de los datos a través de gráficos y visualizaciones. En este sentido, este informe se centrará en mostrar y explicar los resultados de las gráficas obtenidas para cada variable en nuestra base de datos. Estas gráficas proporcionarán una representación visual de los patrones, tendencias y relaciones presentes en los datos.

A medida que avanzamos en el informe, se proporcionará una explicación detallada de cada gráfica, destacando los aspectos más relevantes y significativos. Esto permitirá una comprensión profunda de los datos y ayudará en la interpretación de los hallazgos.

Explicación de lo que se realizó en el código

Se llevó a cabo un análisis univariado de las variables categóricas en la base de datos. Este análisis se centra en estudiar cada variable categórica de forma individual, con el propósito de comprender su distribución y frecuencia en el conjunto de datos. Como se muestra en la *Imagen 1.*

```
[10] #Obtenemos un análisis univariado de las vvariables categóricas
     freq_tbl(Micro_Retailer)
```

	record_id	frequency	percentage
0	dff2998e-af74-4de6-8efd-488aca24e67b	1	0.005848
1	53c25f08-1c1b-4c1c-97d5-b45c940735cc	1	0.005848
2	af5c48b5-a916-47a6-aacc-1128ea1e728e	1	0.005848
3	4790411b-df1d-44f0-a659-6d99e26ae765	1	0.005848
4	e70db20a-25db-412b-9a71-7df285c92a3d	1	0.005848
5	6311502b-fa7b-45fc-817f-c229447b7fe1	1	0.005848
6	ea013169-99dd-4027-ac70-91c66161a1a5	1	0.005848
7	fa403799-e3d6-427b-accc-4160832a263b	1	0.005848
8	b2783738-b018-49e2-849d-da909ee5058a	1	0.005848
9	9c4c7985-fc20-4d2a-ae51-a65c584f14eb	1	0.005848
10	c67ce802-7010-4d33-950d-48902e5c7de2	1	0.005848
11	67df942e-bcee-47c2-aee4-bb51d2d741c1	1	0.005848
12	adadd94c-68ad-4bfa-a816-ba4c46a8a781	1	0.005848
13	93e02631-440e-466f-b05d-69e301fbaab0	1	0.005848
14	0591b725-62b4-4871-b3c5-ccbff36ca577	1	0.005848
15	0a9f9129-7467-4669-957c-722a26bc27cb	1	0.005848
16	2c6652da-1fa7-4b92-ad81-893fbd866a3e	1	0.005848
17	53d2492a-162f-4a91-afa3-764efdd719a1	1	0.005848
18	555a5830-66b1-47b5-b174-5480c3b38f8f	1	0.005848
19	c5097507-8988-4109-8e81-564c57f59382	1	0.005848
20	0f1de586-cf6f-4e28-924d-e85f3024587c	1	0.005848
21	86c15891-fffa-4a40-b7e7-93c88034c245	1	0.005848

Imagen1

Además del análisis univariado de las variables categóricas, se llevó a cabo un proceso de filtro por columnas para determinar qué variables serían objeto de análisis más detallados. Este filtro fue esencial para seleccionar las columnas más relevantes y significativas en función de los objetivos de la investigación. Ver *Imagen 2.*

```
#Filtro por columna
dfcategorias=Micro_Retailer.iloc[ :,[9,15,23,24,26,27,34,35,36,48,53,66,67,68,79,82,83,86,88,95]]
```

Imagen 2

Mediante programación, como se puede observar en la *Imagen 3* E imagen *4,* se implementaron procedimientos para identificar y verificar la existencia de estos valores nulos en las variables de interés. Una vez identificados, se adoptó la estrategia de sustituir estos valores por la categoría 'Sin respuesta'. Esta decisión se basó en la necesidad de mantener la integridad de los datos y garantizar que todas las observaciones se incluyeran en el análisis, incluso cuando los encuestados no proporcionaron una respuesta específica.

```
[13] #Corroboramos valores nulos
     valores_nulos=dfcategorias.isnull().sum()
     valores_nulos

232_type_of_store                          0
184_store_devices                          2
5_change_store_space_last_year            72
6_change_employees_average_salary_last_year  86
49_inventory_records                      61
18_sales_records                          61
186_internet_connection                   63
210_sales_channels                        89
189_payment_methods                       64
33_credit_to_customers                    64
193_sales_planning_tools                  89
311_topups                               114
312_payment_of_utilities                 105
313_home_deliveries                      113
185_place_orders_suppliers               100
192_procurement_planning_tools           103
277_payment_method_suppliers             103
157_frequency_organize_shelves            75
161_actions_stockouts                     79
```

Imagen 3

```
#Eliminamos datos nulos
data_clean_dfcategorias=dfcategorias.copy()
data_clean_dfcategorias=dfcategorias.fillna("SIN RESPUESTA")
data_clean_dfcategorias
```

	232_type_of_store	184_store_devices	5_change_store_space_last_year	6_change_employees_average_salary_last_year	49_inventory_records	18_sale
0	Tailor	POS system (i.e. computer + bar code scanner +...	No cambió	No cambió	Software especializado + computadora	espe cc
1	Grocery store (aka. nanostore)	Dataphone (i.e. card payments)	Aumentó 15%	No cambió	No lo hago	
2	Grocery store (aka. nanostore)	Dataphone (i.e. card payments),POS system (i.e.	No cambió	No cambió	En Excel	espe cc

Imagen 4

Como parte de las medidas de calidad de datos y para asegurarnos de que los datos estuvieran limpios y completos, se realizó una segunda verificación para confirmar que no hubiera datos nulos en el conjunto de datos después de la sustitución previamente mencionada. Ver *Imagen 5.*

```
#Comprobamos datos nulos
valores_nulos=data_clean_dfcategorias.isnull().sum()
valores_nulos

232_type_of_store                          0
184_store_devices                          0
5_change_store_space_last_year             0
6_change_employees_average_salary_last_year  0
49_inventory_records                       0
18_sales_records                           0
186_internet_connection                    0
210_sales_channels                         0
189_payment_methods                        0
33_credit_to_customers                     0
193_sales_planning_tools                   0
311_topups                                 0
312_payment_of_utilities                   0
313_home_deliveries                        0
185_place_orders_suppliers                 0
192_procurement_planning_tools             0
```

Imagen 5

Explicación de gráficas

La primera variable analizada fue la variable “232_type_of_store”. Con la gráfica que se muestra en la *Imagen 6* se logra visualizar que los tipos de tienda que tuvieron mayor presencia en la escuela es de Micro restaurantes y Tienda de comestibles

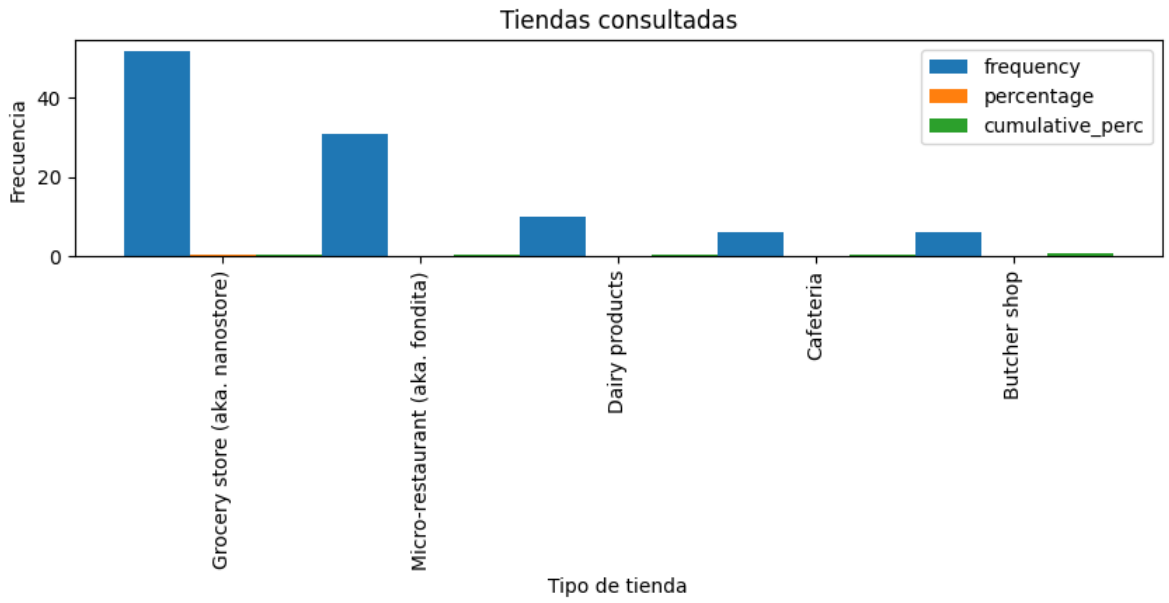

Imagen 6

La segunda variable analizada fue ” 184_store_devices”. Con la gráfica que se muestra en la *Imagen 7* de concluye que la mayoría de las tiendas consultadas ocupan un télefono inteligente para el almacenamiento de sus productos

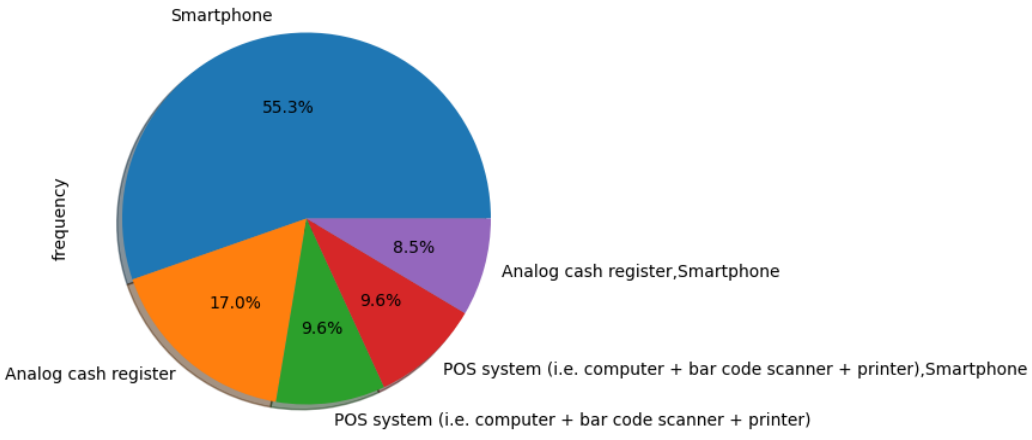

Imagen 7

La tercera variable analizada fue ”5_change_store_space_last_year”. Con esta gráfica que se muestra en la *Imagen 8* podemos observar un resultado parejo donde las tiendas encuestadas no tuvieron un cambio en el espacio de la tienda el último año, sin embargo, nos encontramos con un alto valor en el número de "sin respuesta" por parte de los encuestados. Pocas tiendas encuestadas tuvieron cambios en los espacios de su tienda el último año.

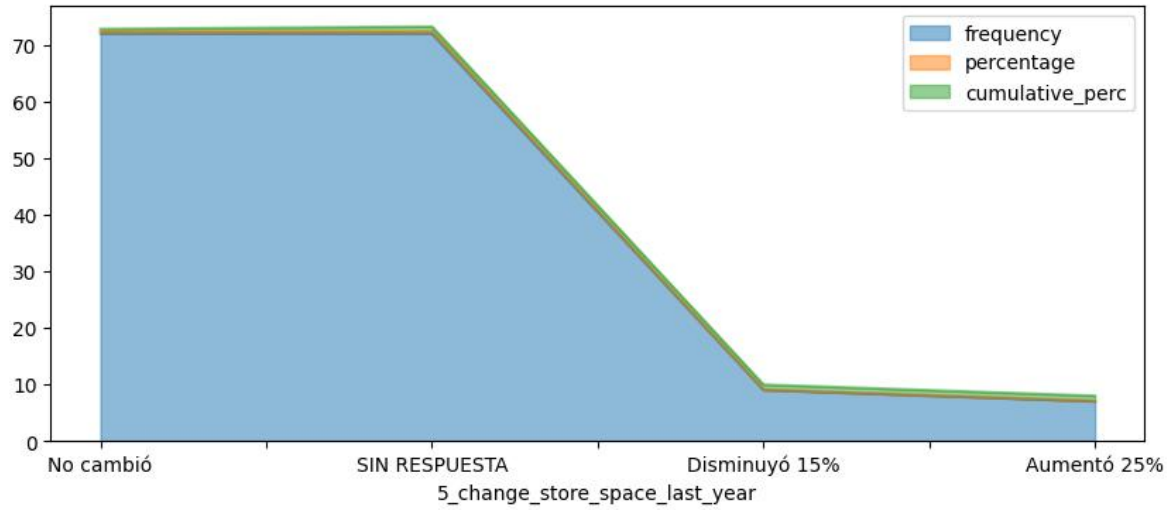

Imagen 8

La cuarta variable analizada fue “6_change_employees_average_salary_last_year”. En esta variable podemos notar que la mayoría de los encuestados no dieron una respuesta a que si había habido un cambio en el promedio de salarios para los empleados, pero la segunda respuesta más alta es que no cambiaron los salarios y solamente el 9.3% de los encuestados tuvieron un aumento de 15%. Ver *Imagen 9*

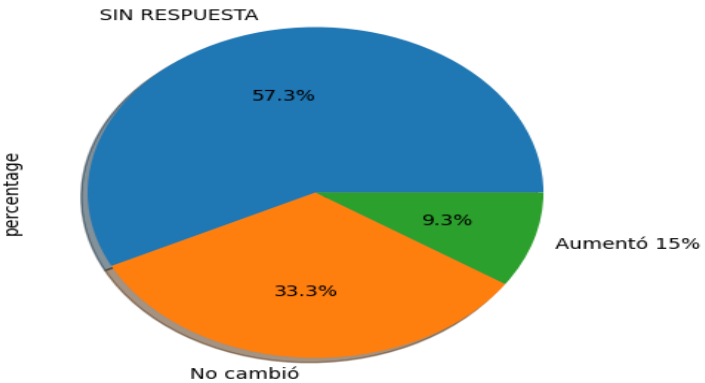

Imagen 9

La quinta variable analizada fue “49_inventory_records”. En la *Imagen 10* se puede observar que la mayoría de los encuestados no dieron una respuesta sobre donde hacen los registros de inventario, sin embargo, se nota que hay un valor significativo en donde los encuestados hacen sus registros de inventario en papel.

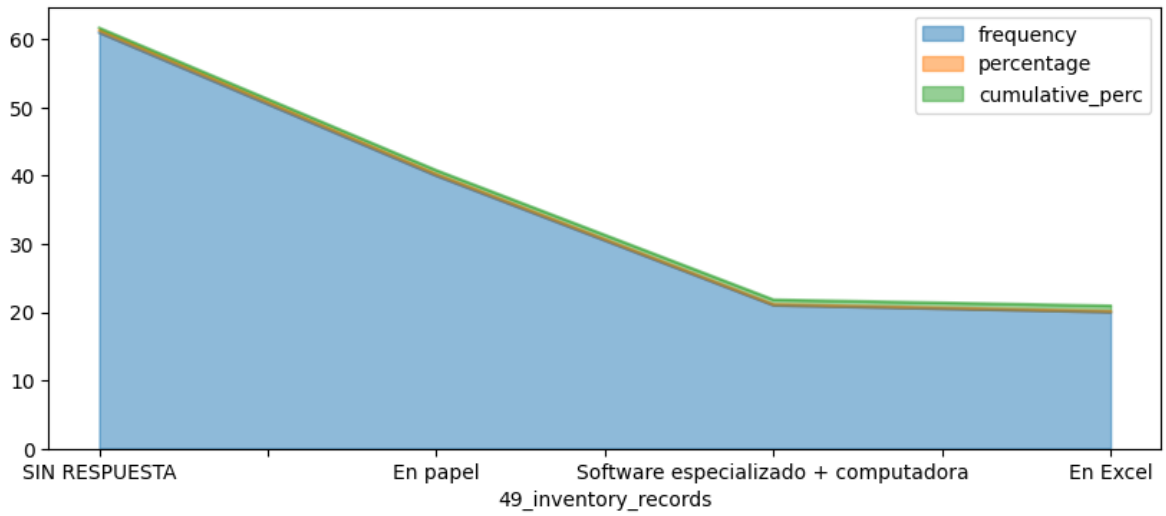

Imagen 10

La sexta variable analizada fue “18_sales_records”. En la *Imagen 11* podemos seguir observando que una gran parte de los encuestados hacen el registro de sus ventas a papel, y una minoría lo hace en Excel y/o computadora. En esta variable se observó una nueva respuesta mostrada en los resultados que muestran que el 9% no hacen registro de sus ventas

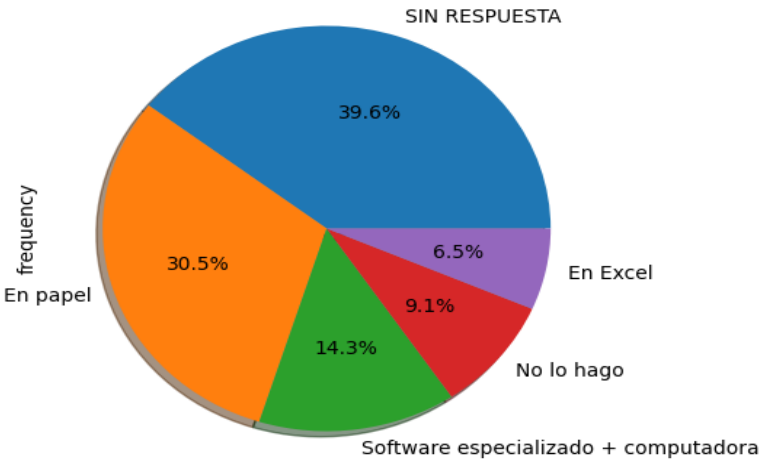

Imagen 11

La séptima variable analizada fue “186_internet_connection”. En la *Imagen 12* se puede observar que la mayoría de las tiendas sí cuentan con conexión a internet, una menor parte no cuenta con este elemento. Sigue presente de manera significante que hubo encuestados que no contestaron esta pregunta.

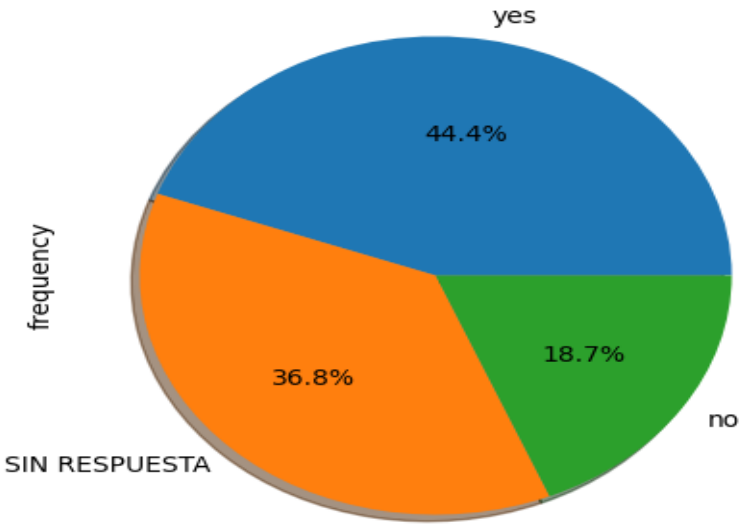

Imagen 12

La octava variable analizada fue “210_sales_channels”. Nos encontramos con una variable que tampoco hay mucha respuesta por parte de los encuestados, pero se puede observar en las *Imagen 13* que los encuestados que respondieron la pregunta ocupan redes sociales para realizar sus canales de ventas.

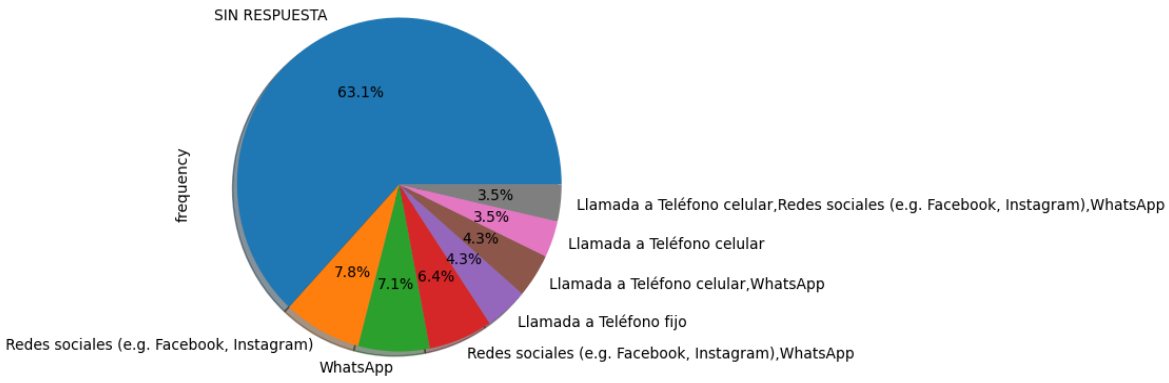

Imagen 13

La novena variable analizada fue “189_payment_methods”. Se puede observar en la *Imagen 14* que por parte de los encuestados que la mayoría tienen como método principal el pago en efectivo.

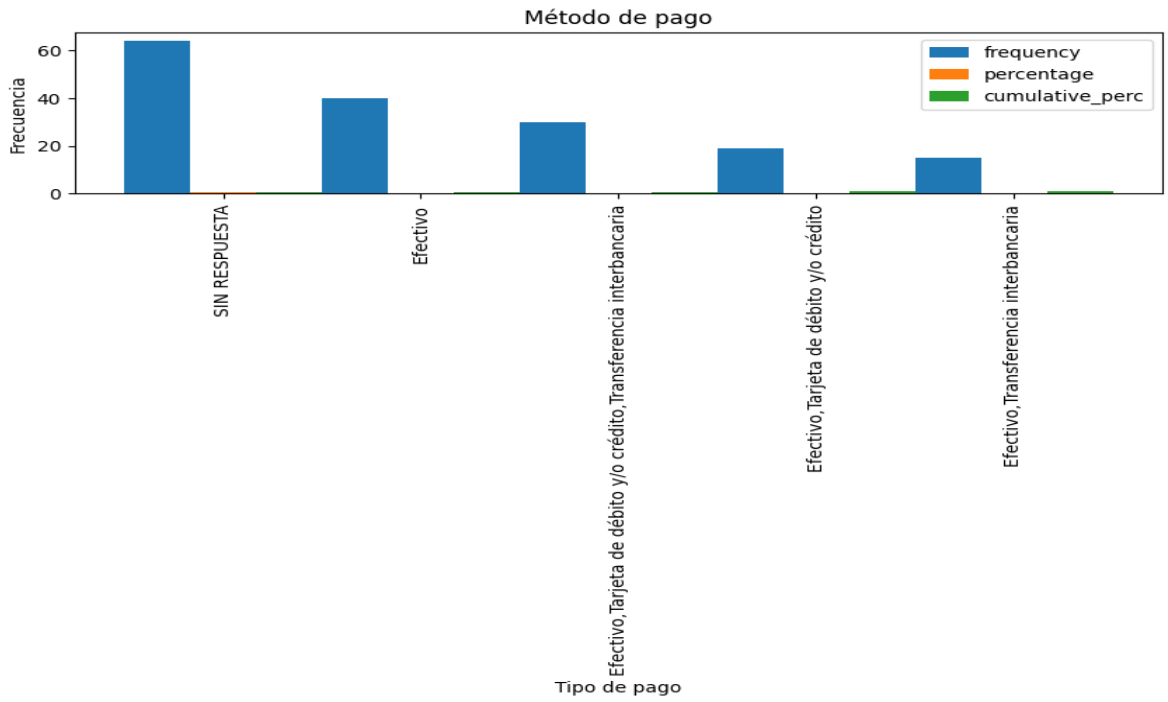

Imagen 14

La décima variable analizada fue “33_credit_to_customers”. En esta variable categórica se puede visualizar a través de las diferentes gráficas que la mayoría de las tiendas no ofrecen créditos a sus clientes, como se muestra en la *Imagen 15*.

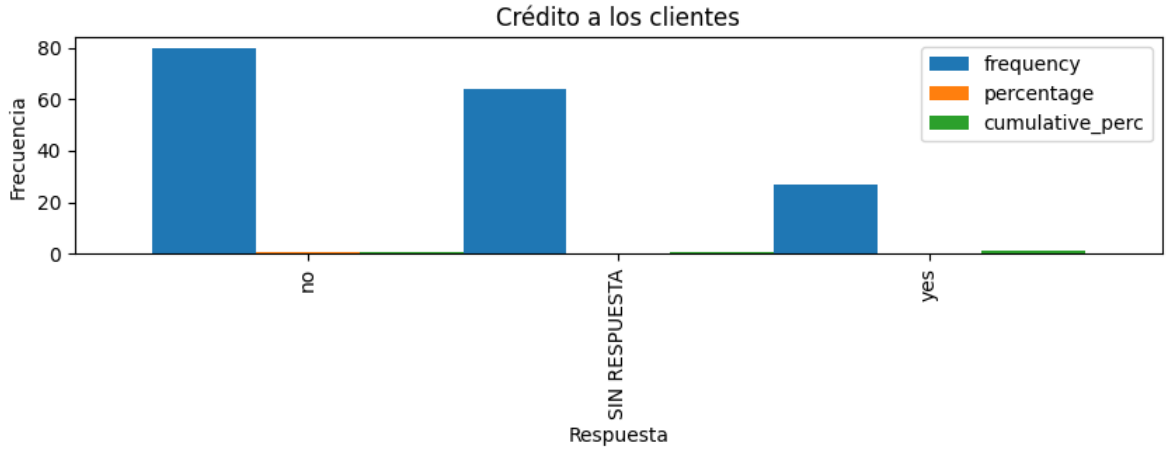

Imagen 15

La undécima variable analizada fue “193_sales_planning_tools”. En esta variable, como se muestra en la *Imagen 16,* se puede identificar nuevamente que de los encuestados la mayoría no respondió esta pregunta, pero muestra como otras personas encargadas de la tienda no hacen planeación de ventas o si es que lo hacen, lo hacen en cuaderno o en software especializado.

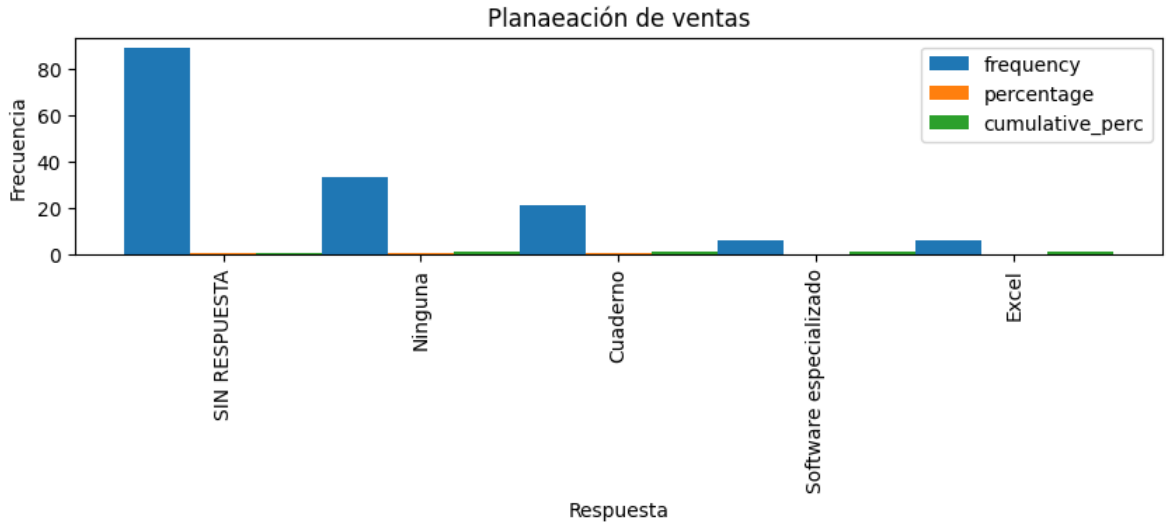

Imagen 16

La duodécima variable analizada fue “311_topups”. De las personas que respondieron a esta pregunta se nota la diferencia significativa en donde los encuestados están totalmente en desacuerdo de los top ups en las tiendas, se muestran en la *Imagen 17.*

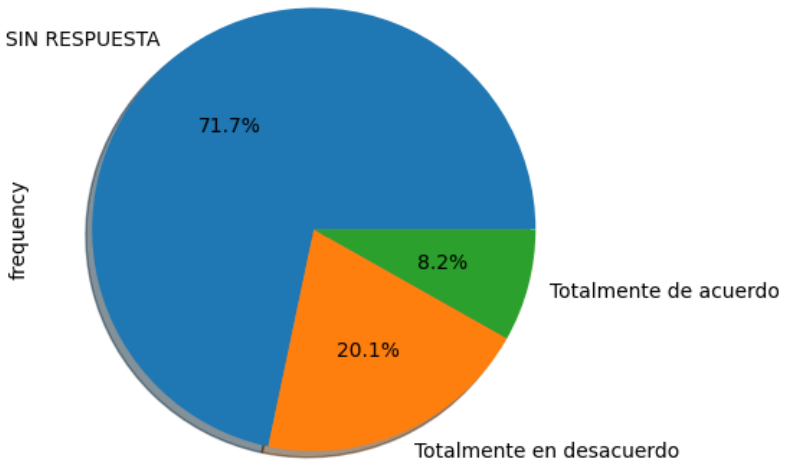

Imagen 17

La decimotercera variable analizada fue “312_payment_of_utilities”. En la *Imagen 18*, se muestra que de las personas que respondieron esta pregunta se logra apreciar que hay una cantidad mayor de personas que están totalmente en desacuerdo del pago por utilidades que las personas que están totalmente de acuerdo

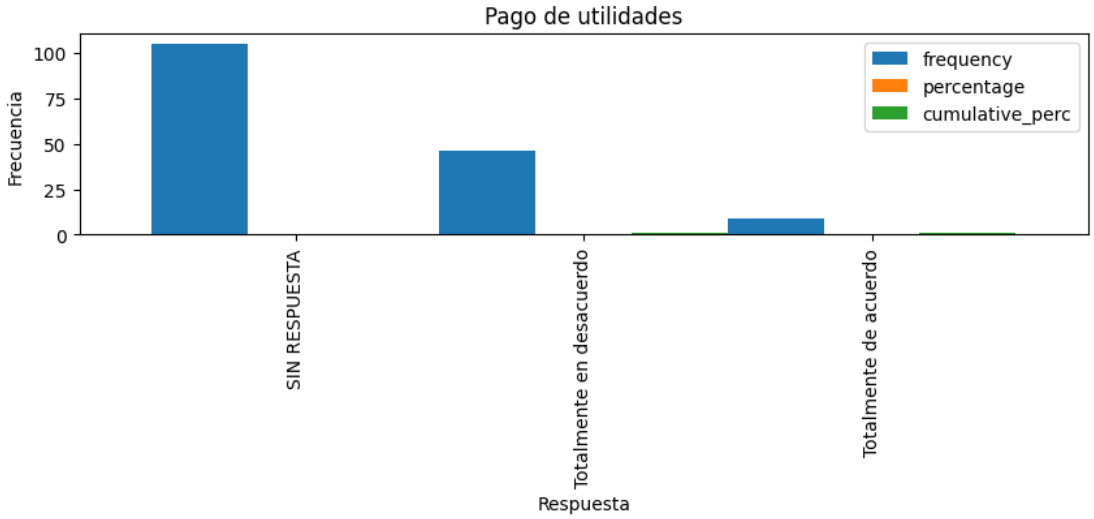

Imagen 18

La decimocuarta variable analizada fue “313_home_deliveries”. De las personas que contestaron esta pregunta se logra apreciar en la *Imagen 19* que las personas que están totalmente en desacuerdo es mayor que las personas que están en totalmente de acuerdo para las entrega a domicilio.

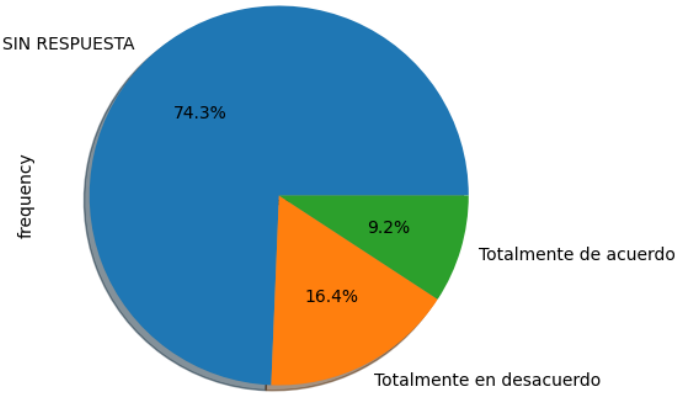

Imagen 19

La decimoquinta variable analizada fue “185_place_orders_suppliers”. En la *Imagen 20*, de las personas que contestaron esta pregunta se puede apreciar que hay mayor cantidad que hacen sus pedidos a proveedores en persona que a través de mensajería instantánea o llamada.

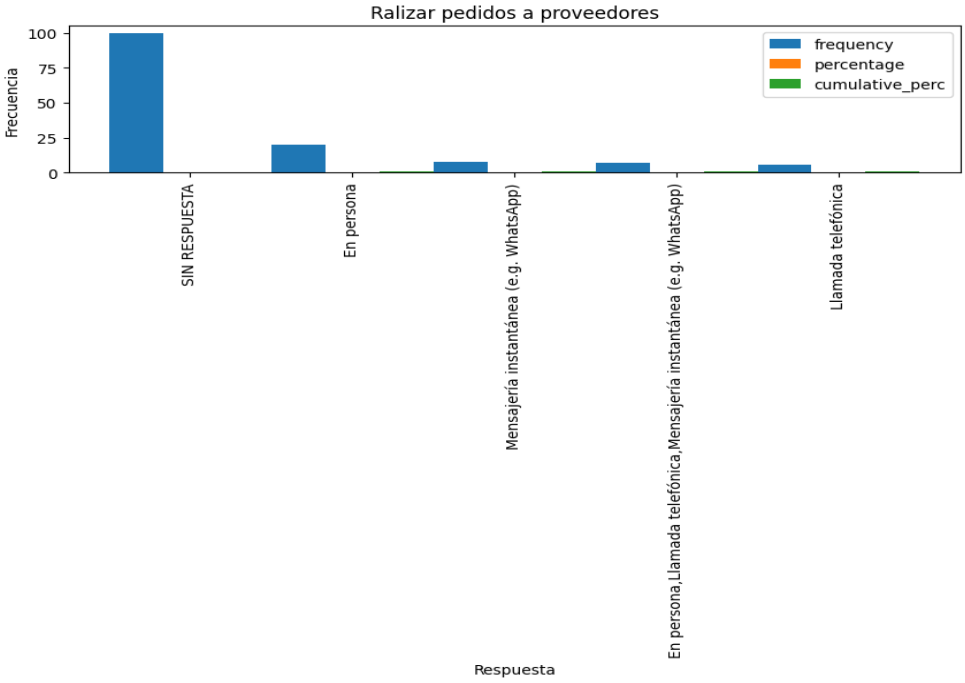

Imagen 20

La decimosexta variable analizada fue “192_procurement_planning_tools”. En esta variable, de las personas que contestaron la encuesta se puede visualizar en la *Imagen 21* que casi hay un empate en las personas que ocupan herramientas de planificación de adquisiciones entre cuaderno y los que no ocupan ninguna herramienta, aunque es un valor más grande el primero de estos dos. Y la minoría es que ocupan la herramienta Excel para esta actividad

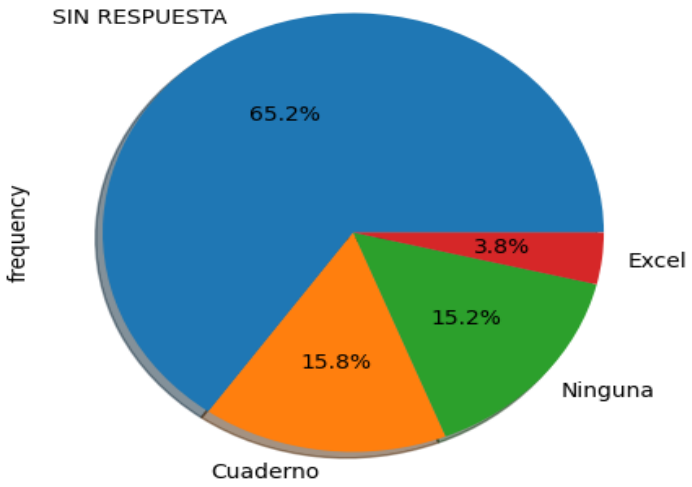

Imagen 21

La decimoséptima variable analizada fue "277_payment_method_suppliers". En esta variable se puede visualizar en la *Imagen 22* que de las personas que contestaron esta pregunta, el método de pago de proveedores que más se utiliza es únicamente en efectivo, con un porcentaje menor se encuentran ambos métodos (efectivo y transferencia) y por último y más pequeño porcentaje solo transferencia interbancaria

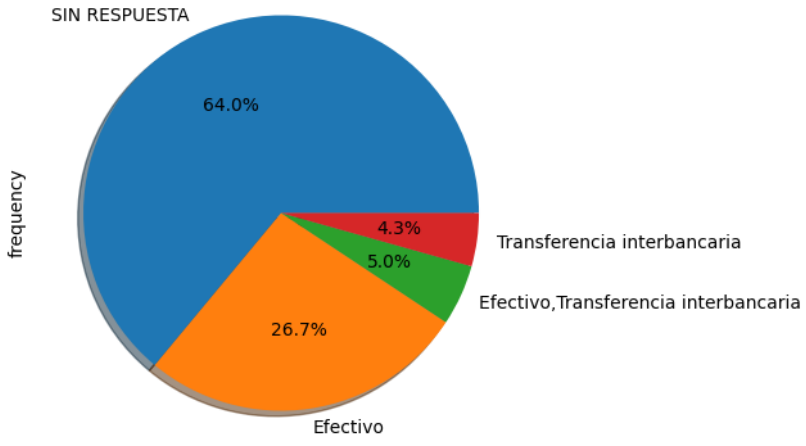

Imagen 22

La decimoctava variable analizada fue "157_frequency_organize_shelves". De las personas que contestaron esta encuesta se puede visualizaren la *Imagen 23* que la mayoría de las personas

organizan sus estantes de manera semanal o diariamente, por último, con valores más pequeños son las personas que nunca organizan sus estantes o lo realizan de manera mensual

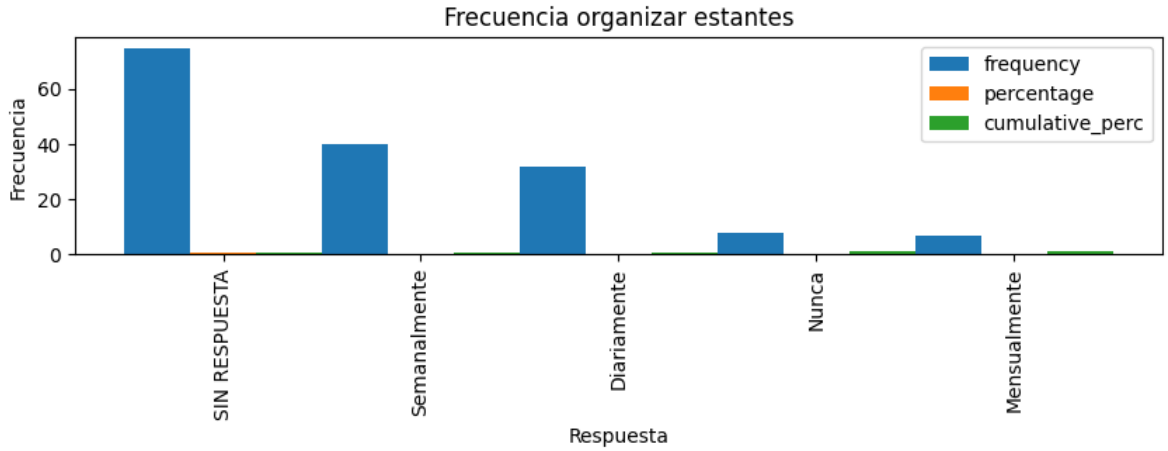

Imagen 23

La decimonovena variable analizada fue "161_actions_stockouts". En esta variable se puede visualizar en la *Imagen 24* que de las personas que respondieron esta pregunta, la mayoría de las personas en sus tiendas cuando se agotan los productos solicitan tiempo al cliente para conseguir el producto agotado, después con un valor menor las personas ofrecen un producto sustituto y el valor más pequeño hacen ambas acciones, ofrecen un producto sustituto o solicitan tiempo al cliente para conseguir el producto agotado.

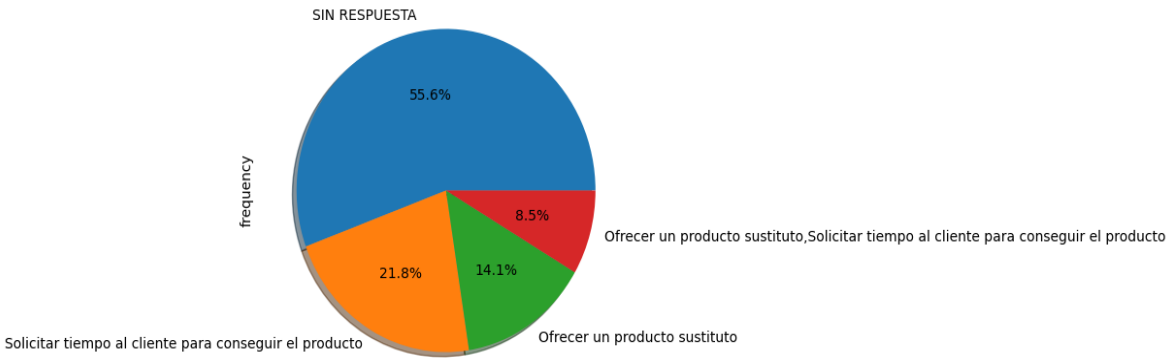

Imagen 24

La vigésima y última variable analizada fue "260_technology_scalable". En esta variable se puede visualizar en la *Imagen 25* que de las personas que contestaron la pregunta, tuvo mayor cantidad la

respuesta "totalmente de acuerdo" y "de acuerdo" a la variable de tecnología escalable para sus negocios, sin embargo, con un valor menor también se observa que hubo una cantidad considerable de las personas que están "en desacuerdo" o "totalmente en desacuerdo".

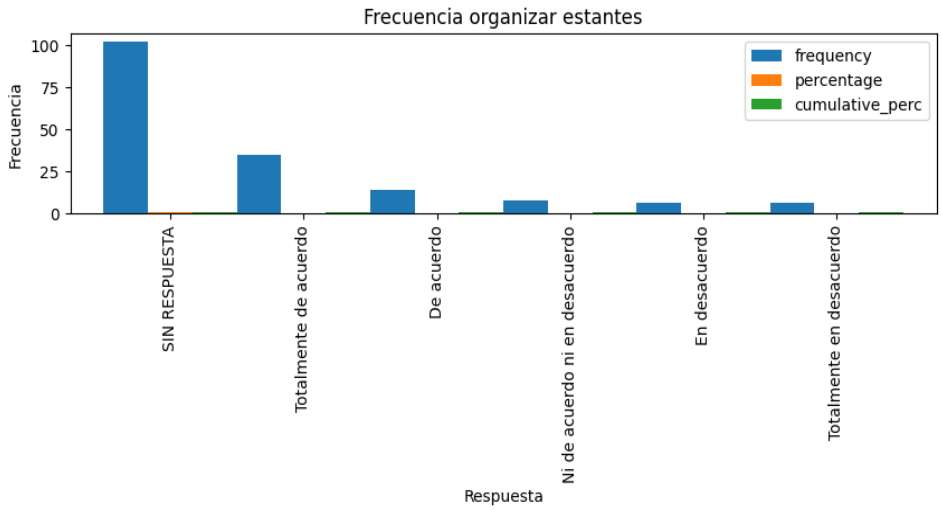

Imagen 25

Conclusión

En este reporte, se realizó un análisis completo de datos de tiendas, abordando la limpieza de datos, la detección de valores atípicos y la selección de variables clave. Se aplicaron técnicas efectivas para asegurar la calidad de los datos y se generaron visualizaciones informativas. Este proceso proporcionó una comprensión sólida de la información de la encuesta. En resumen, el análisis de datos desempeña un papel crucial en la toma de decisiones basadas en datos y la comprensión de los negocios