Tecnológico de Monterrey

Analítica de datos y herramientas de inteligencia artificial II TI3001C

Actividad 4

Extracción de características

Estefanía López Ponce A01654214

Septiembre 28, 2023

Descripción del Set de Datos

El conjunto de datos denominado "microretailer_mit_lift_lab.xlsx" que se nos ha proporcionado consta de un total de 171 filas, numeradas desde 0 hasta 170. Este conjunto de datos posee un total de 106 columnas, dentro de las cuales se encuentran 23 columnas con datos de tipo float, 2 columnas con datos de tipo entero y 81 columnas con datos de tipo objeto.

Exploración del Set de Datos en Python

El procedimiento inicial consiste en adquirir un conocimiento integral de la composición del conjunto de datos mediante la ejecución del comando `info()`.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 171 entries, 0 to 170
Data columns (total 106 columns):
 #    Column                                                Dtype
---   ------                                                -----
 0    _record_id                                            object
 1    _title                                                object
 2    _server_updated_at                                    object
 3    _updated_by                                           object
 4    _geometry                                             object
 5    _latitude                                             float64
 6    _longitude                                            float64
 7    228_store_name                                        object
 8    229_store_picture                                     object
 9    232_type_of_store                                     object
 10   108_does_the_micro_retailer_has_a_barred_window_      object
 11   99_does_the_micro_retailer_exhibits_products_outside_ object
 12   102_does_the_micro_retailer_sells_fresh_products_     object
 13   97_number_of_customers_in_store                       int64
 14   268_number_fridges                                    float64
 15   184_store_devices                                     object
 16   104_how_many_shelves_does_the_micro_retailer_have     int64
 17   pictures_of_shelves_if_possible                       object
 18   hola_somos_estudiantes_del_tec_de_monterrey_estamos_realiz object
 19   le_comento_que_toda_esta_informacin_es_confidencial   object
 20   233_date_establishment                                object
 21   2_current_permanent_employees                         float64
 22   4_number_permanent_employees_last_year                float64
 23   5_change_store_space_last_year                        object
 24   6_change_employees_average_salary_last_year           object
 25   20_reviews_finances_monthly                           object
 26   49_inventory_records                                  object
 27   18_sales_records                                      object
 28   155_sales_registers_used_for                          object
 29   103_number_own_fridges                                float64
 30   19_tax_id                                             object
 31   145_number_direct_competitors                         float64
 32   310_burnout                                           object
 33   24_burnout                                            float64
```

Figura 1. Informacion del conjunto de datos.

Con el fin de efectuar un análisis visual de determinadas variables, se ha procedido a la extracción de un subconjunto de columnas con el propósito de conformar un nuevo DataFrame denominado "Data20". Este nuevo DataFrame albergará un total de 20 columnas, todas ellas de tipo objeto, con el propósito de facilitar una clasificación más efectiva de los datos.

```
# Seleccionar las 20 columnas deseadas para analizar
column_indices = [3, 9, 24, 26, 27, 34, 35, 36, 37, 38, 47, 50, 53, 55, 61, 62, 68, 79, 80, 83]
data20 = data.iloc[:, column_indices]
data20
```

Figura 2. Generación del DataFrame a evaluar.

A continuación, se procede a la verificación de la presencia de datos nulos en el DataFrame utilizando el comando "isnull().sum()". Este proceso nos proporcionará la cantidad de datos faltantes por columna.

```
_updated_by                                  0
232_type_of_store                            0
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
28_prefered_payment_method                   169
31_different_prices_payment_method           164
214_customer_relationship_tools              76
35_interest_rate                             150
193_sales_planning_tools                     89
60_services                                  94
77_delivery_timeslots                        169
79_delivery_vehicle                          169
313_home_deliveries                          113
185_place_orders_suppliers                   100
53_special_conditions_suppliers              124
277_payment_method_suppliers                 103
dtype: int64
```

Figura 3. Datos faltantes por columnas.

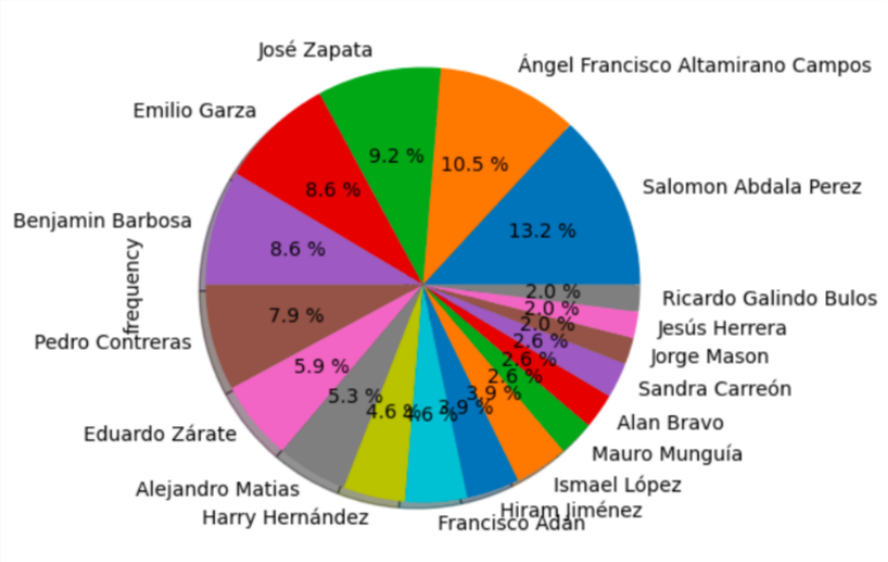
Con el objetivo de abordar la cuestión de los datos faltantes en estas columnas, se ha creado un nuevo DataFrame denominado "data20_filled". Esto se llevó a cabo con la finalidad de llevar a cabo un análisis más preciso. En este proceso, se procedió a una nueva evaluación para asegurar que el método de relleno de datos fuera efectivo y resultara en un conjunto de datos completo y coherente.

Gráficas

Para cada columna se evalio la frecuencia de repeticion de datos, se realiza un filtro para los valores mas relevantes de las variables categoricas seleccionadas, en este caso se omitian aquellos valores que solo se repitieran 1 o 2 veces, dependiendo de la cantidad de clasificaciones que se encuentren. Después de obtener este filtro, se establece el indice como la columna que se busca graficar, mostrando el indice, la frecuencia, el porcentaje y el porcentaje acumulado.

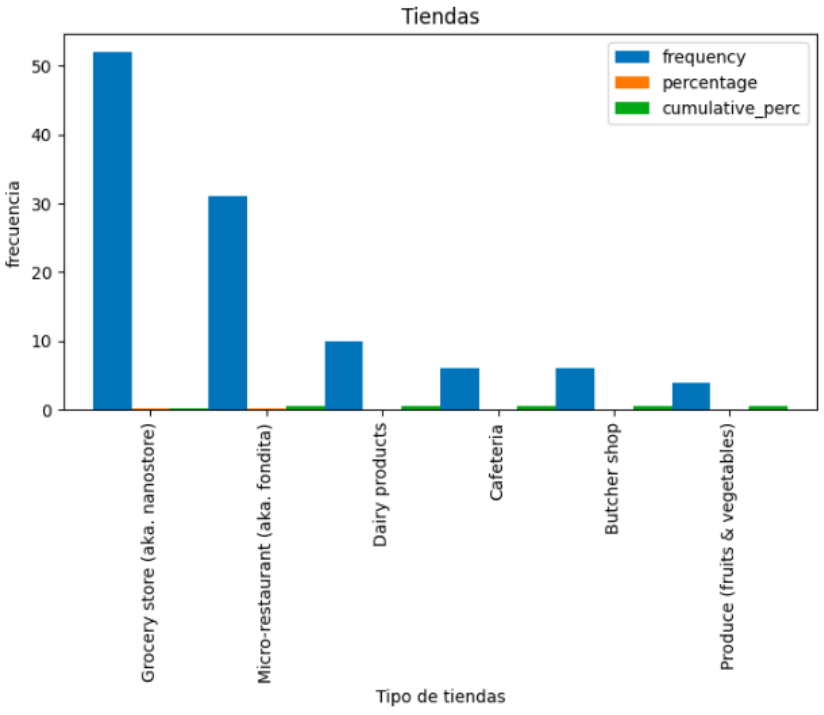
1. **Updated by**

Podemos observar que la gran mayoria de datos fuerón actualizados por Salomon Abdala(20 datos), Angel francisco(16 datos) , Jose Zapata(14 datos), Emili Garza y Benjamin Barbosa(13 datos).



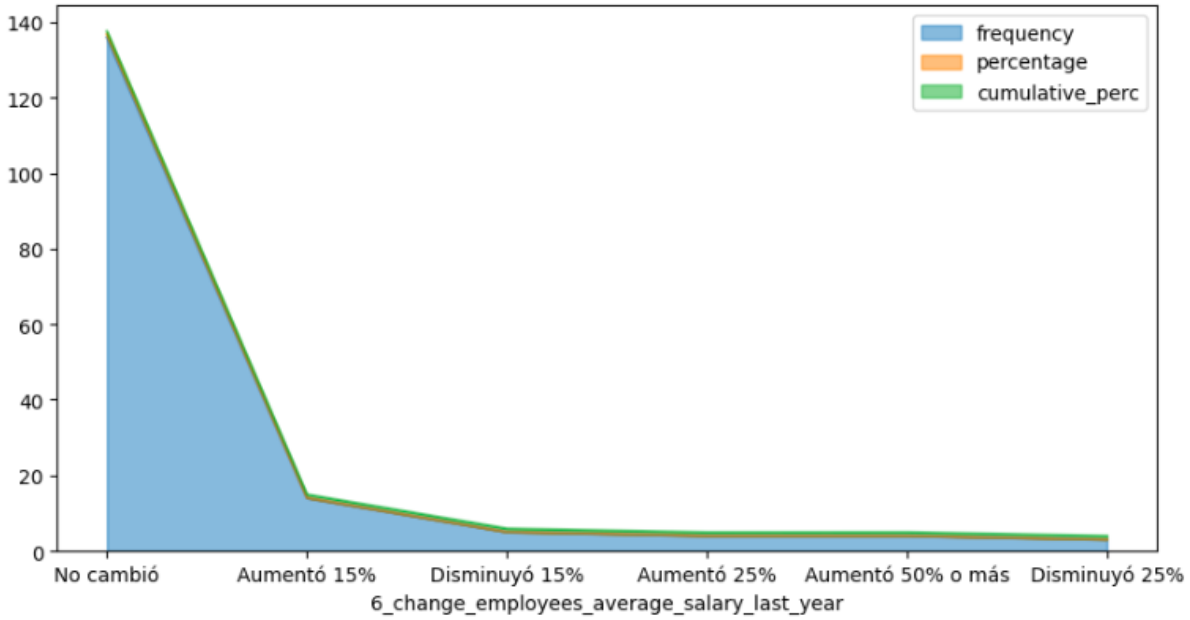
2. **Type of Store**

En la grafica podemos ver que la mayoria de los datos pertenece al tipo Grocery store, seguido de micro restaurantes.



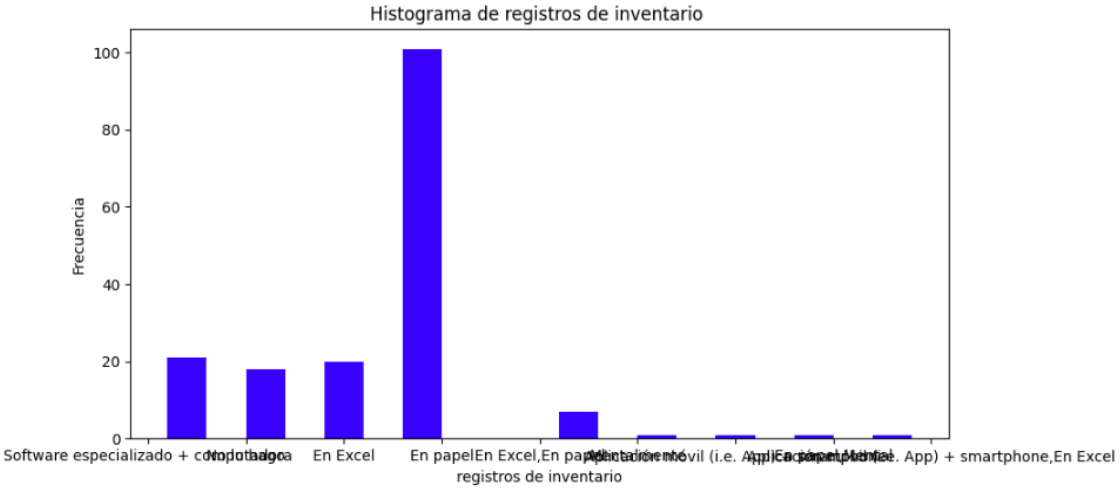
3. **change_employees_average_salary_last_year**

Se observa que en la mayoria de los establecimientos (136) no tuvieron un cambio dentro del promedio del salario para los empleados en el ultimo año, 14 establecimientos lo aumento un 15%, 5 lo disminuyo el 15%, 4 lo aumento el 25%, 4 lo aumento al 50% o más y 3 lo disminuyo el 25%.



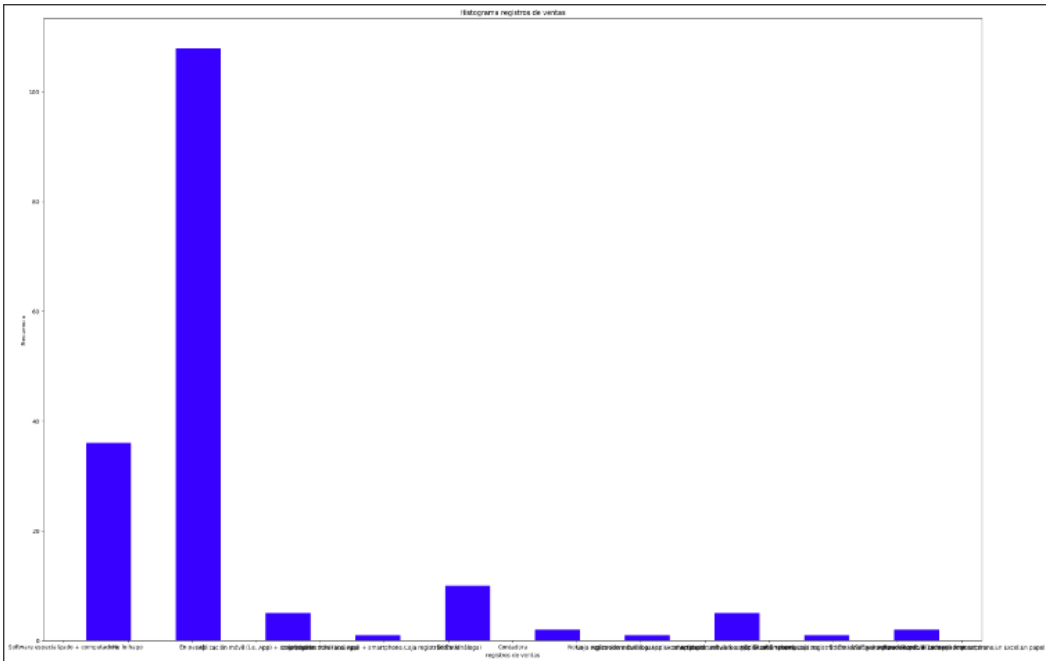
4. **inventory_records**

En el histograma de registro de inventario se visualiza que los establecimientos llevan el seguimeinto de inventario en papel y las otras dos herramientas destacadas son el uso de excel y en un software especializado.



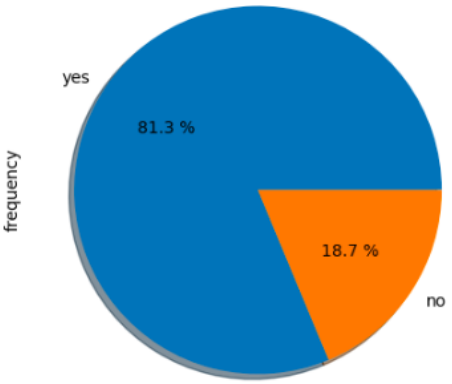
5. **sales_records**

De la misma manera en la que los establecimientos llevan el seguimiento de inventarios, el seguimiento de ventas se lleva a cabo mayormente en papel, un software especializado o simplemente no llevan un registro.



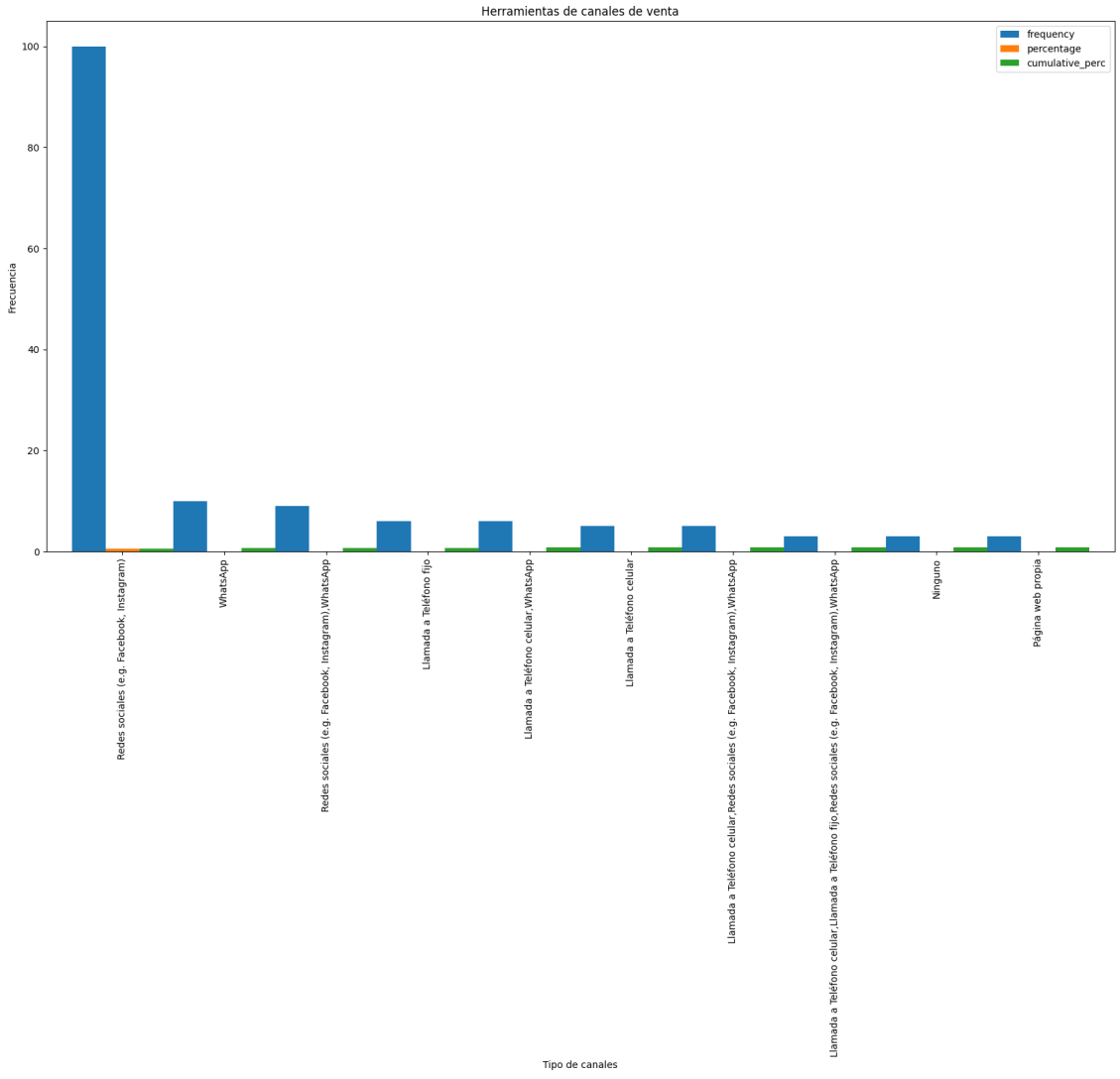
6. **internet_connection**

En la grafica de pastel se muestran las dos categorias que indican si el local cuenta con internet, en este caso el 81.3% de estos tiene el servicio y solo el 18.7% no.



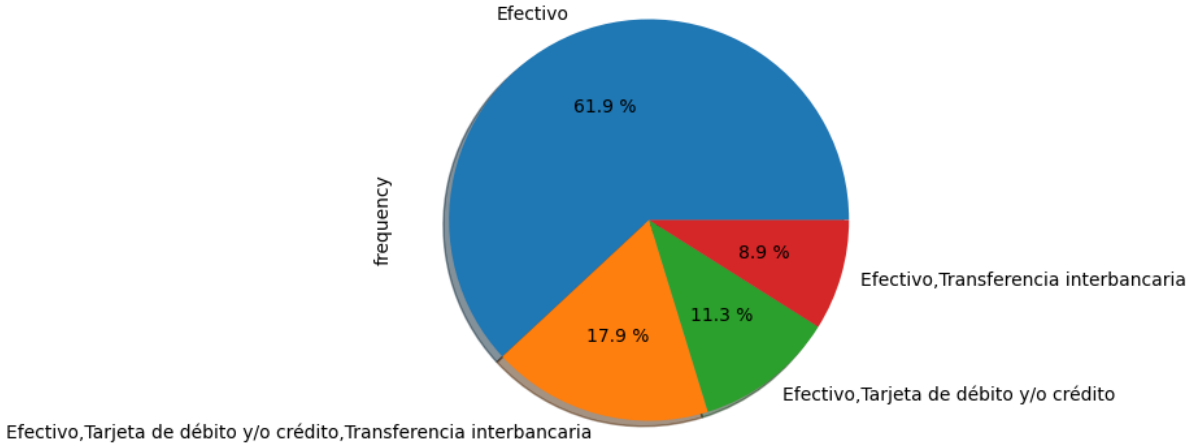
7. sales_channels

La columna contiene los canales de venta que los establecimientos usan para promocionar sus ventas. El canal más usado es el de redes sociales como facebook e instagram y los menos usados son paguida web propia, nungun canal y Llamada a Teléfono celula.



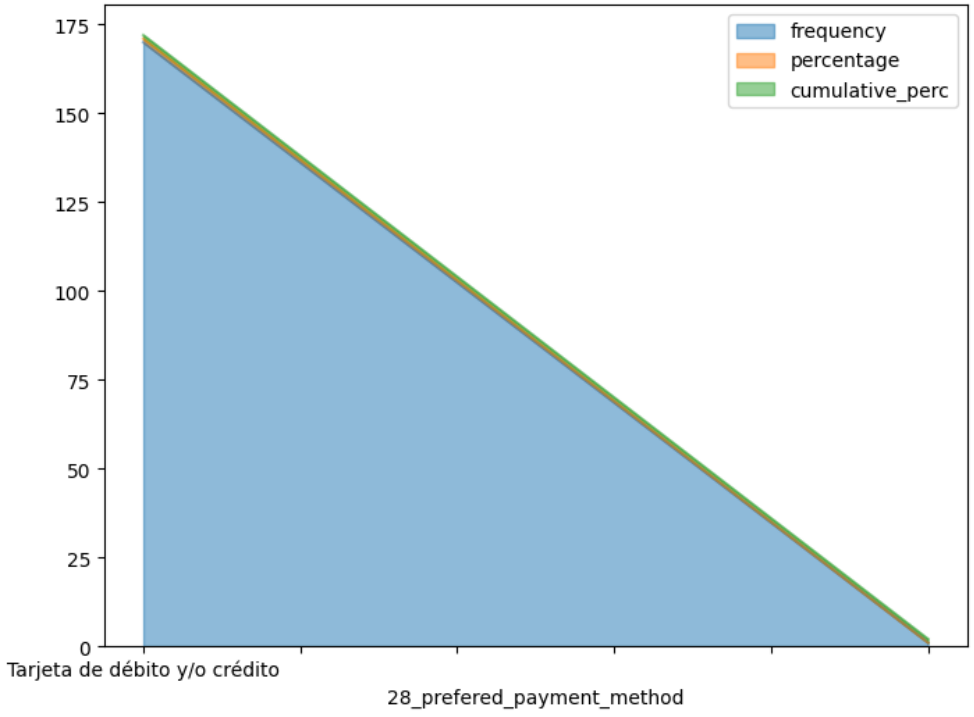
8. payment_methods

El metodo de pago únicamente con efectivo es el que más frecuenta en los establecimientos analizados.



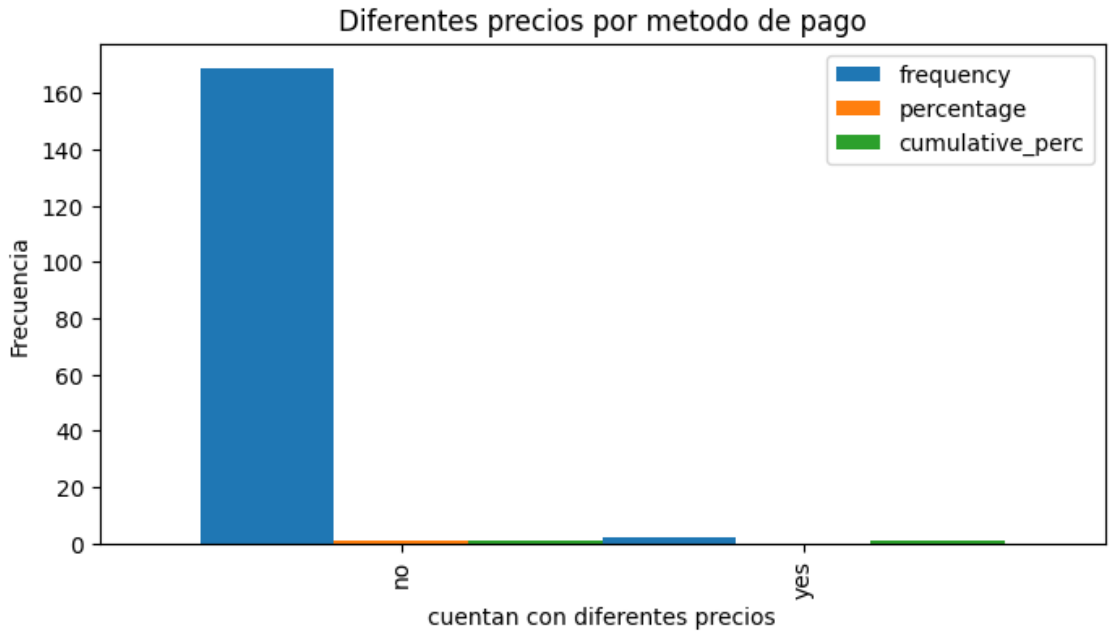
9. prefered_payment_method

El metodo de pago que los clientes prefieren es el de Tarjeta de debito o de credito.



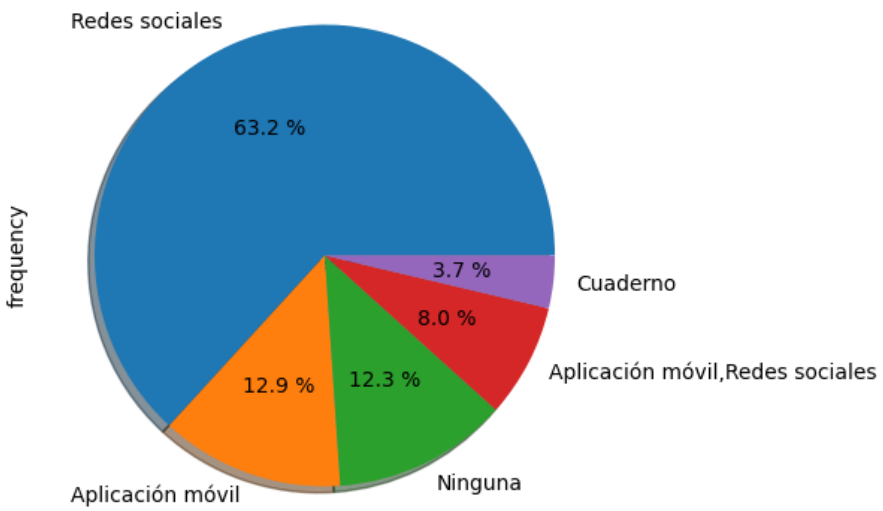
10. different_prices_payment_method

El histograma no muestra si el local aplica un diferente precio si se realiza el pago por diferentes metodos, 169 de ellos no realiza ningun cambio.



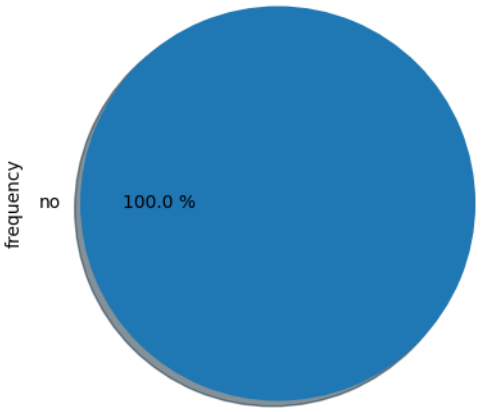
11. customer_relationship_tools

Se identifica las herramientas con las que se mantiene una relación con los clientes, el mayor porcentaje lo obtienen las redes sociales.



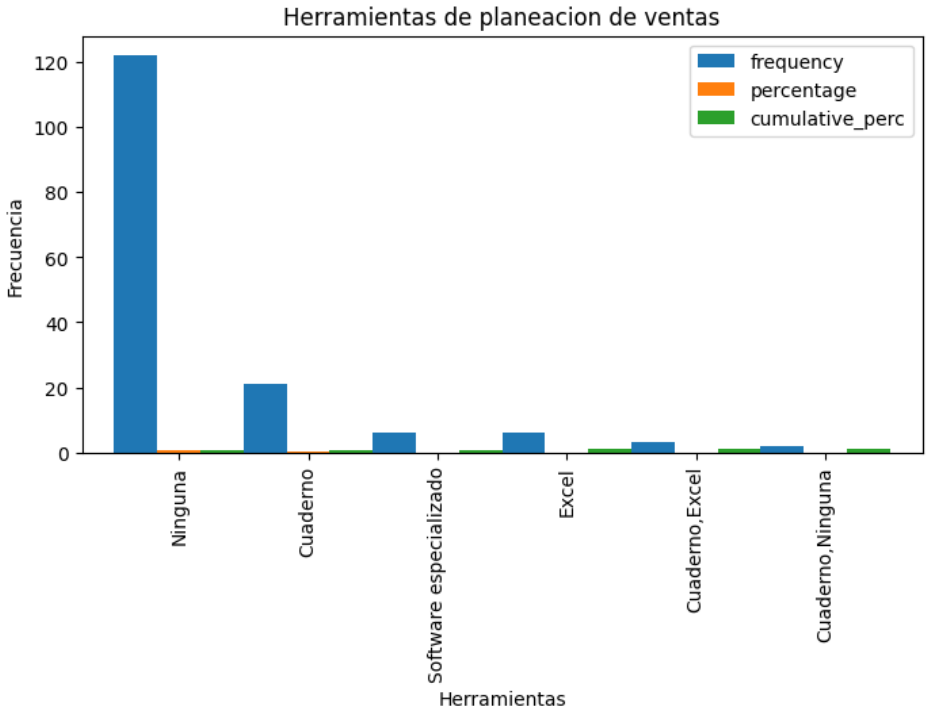
12. interest_rate

Ningun establecimiento aplica una tasa de interes.



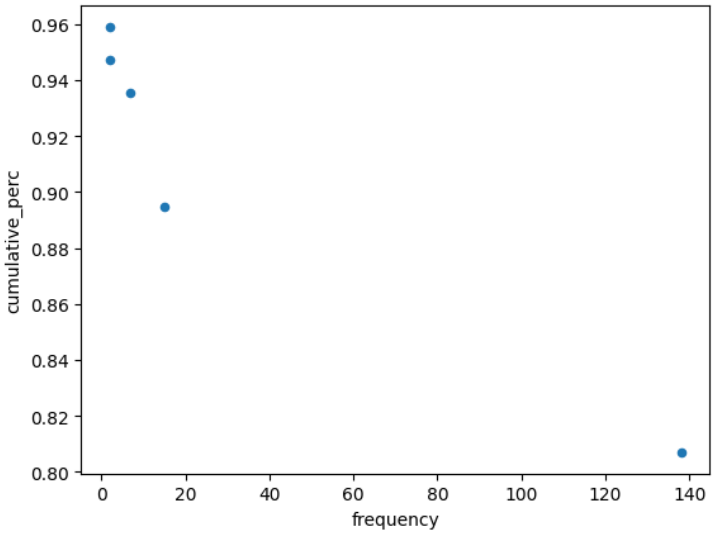
13. sales_planning_tools

La mayor parte de los comercios no hace uso de herramientas para realizar una planificacion de ventas.



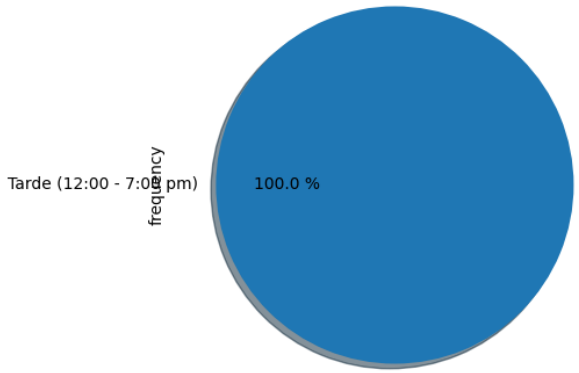
14. services

La siguiente grafica nos muestra la frecuencia en que se repite el mismo servicio que se brinda en el negocio, en este caso el dato menos repetido es servicio de entrega a domicilio, mientras que la mayoria no cuenta con ningun servicio extra.



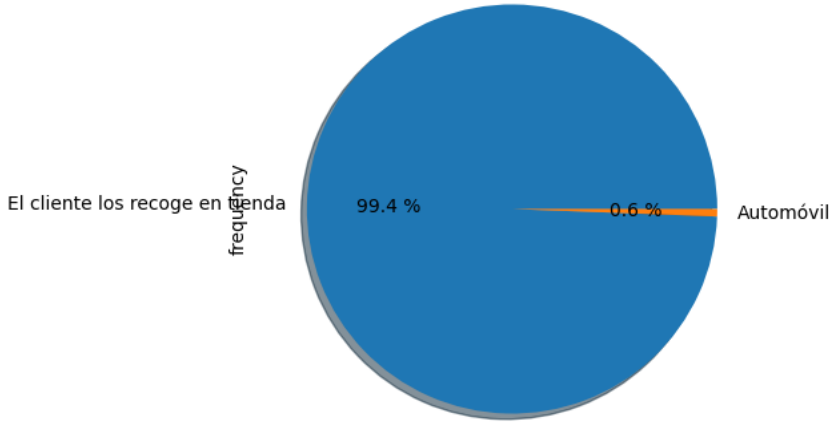
15. delivery_timeslos

Todas las tendas se dedican a hacer entregas durante el horario de la tarde, de las 12:00 a las 7:00pm



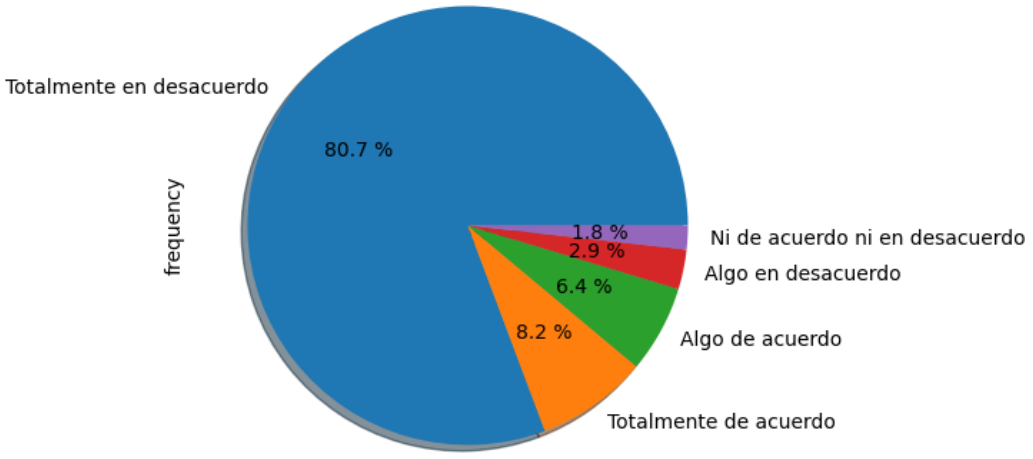
16. delivery_vehicle

El 99.4% no realiza entregas a domicilo por lo que el cliente los recoge en la tienda, solo el 0.6% cuenta con un automovil para realzar estas entregas.



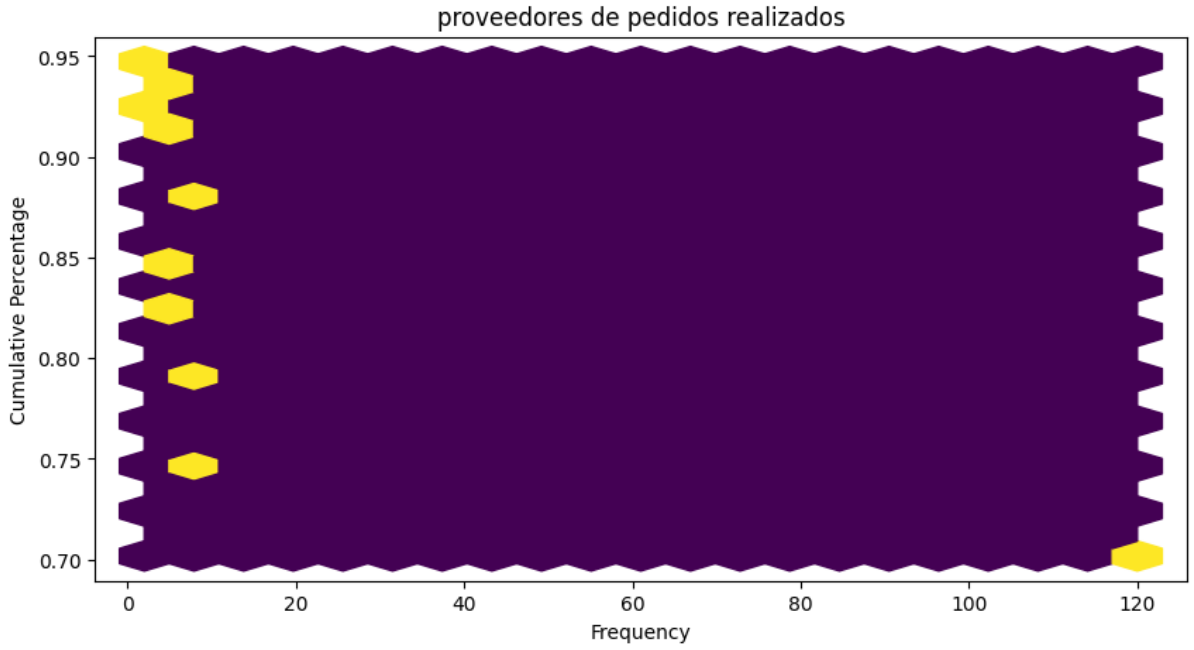
17. home_delivery

Cerca del 81% de los negocios esta en total desacuerdo de realizar entregas a domicilio.



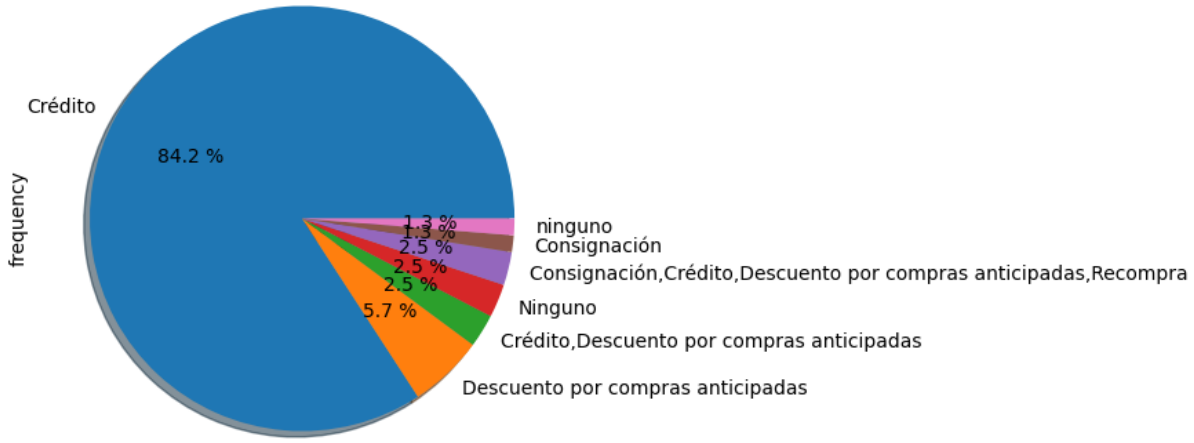
18. place_order_suppliers

La frecuencia con la forma en la que las tiendas realizan los pedidos a los provedores son en persona (120) y por mensajeria instantanea (8)



19. special_conditions_suppliers

El 84% de los negocios tiene como condición especiales para los provedores otorgarles credito.



20. payment_method_suppliers

El metodo de pago mayor mente usado para realizar pagos a los provedores es en efectivo, seguido por la transferencia interbancaria y la tarjeta de credito o debito.