Documento Final Instacart

Pipelines

pipeline_instacartdb:

1. Data Loader: Se encarga de sacar los datos de MySql. Lo hace sacando dfs de cada tabla
2. Data Exporter: Se encarga de exportar los datos hacie el schema RAW. Lo hace enviando los dfs a formato sql

pipeline_instacartsnowf:

1. Data Loader: Se encarga de sacar los datos del schema RAW de Snowflake. Lo hace sacando dfs de cada tabla
2. Transformer: Se encarga de realizar las transformaciones necesarias descritas en el plan de acción. Las realiza en los dfs con pandas
3. Data Exporter: Se encarga de exportar los datos hacie el schema CLEAN. Lo hace enviando los dfs a formato sql

Plan de acción de acuerdo al EDA

1. Limpieza y transformación de datos

- **Valores nulos**:
  - Tabla INSTACART_ORDERS: Para DAYS_SINCE_PRIOR_ORDER, los valores nulos corresponden a la primera compra de los usuarios y llenarlos con 0.
  - Tabla ORDER_PRODUCTS: Para ADD_TO_CART_ORDER, los valores nulos pueden ser eliminados
  - Tabla PRODUCTS: Para PRODUCT_NAME, asignar una etiqueta como Desconocido.
- **Duplicados**:
  - Remover los 15 registros duplicados en INSTACART_ORDERS.

- **Conversión de tipos**:
    - Convertir DAYS_SINCE_PRIOR_ORDER (Tabla INSTACART_ORDERS) y ADD_TO_CART_ORDER (Tabla ORDER_PRODUCTS) a enteros (int)
- **Outliers**:
    - Representan comportamientos válidos de los clientes; por lo que, no se modificaran

Normalización

Si necesitaramos un modelo de machine learning

1. **Tabla:** INSTACART_ORDERS
    - ORDER_HOUR_OF_DAY: Escalar los valores entre 0 y 1.

Modelo Dimensional (Star Schema)

Tablas de Dimensiones:

1. **dim_products**
    - PRODUCT_ID (PK)
    - PRODUCT_NAME
    - AISLE_ID (FK)
    - DEPARTMENT_ID (FK)
2. **dim_aisles**
    - AISLE_ID (PK)
    - AISLE
3. **dim_departments**
    - DEPARTMENT_ID (PK)
    - DEPARTMENT

Tabla de Hechos:

1. **fct_order_products**

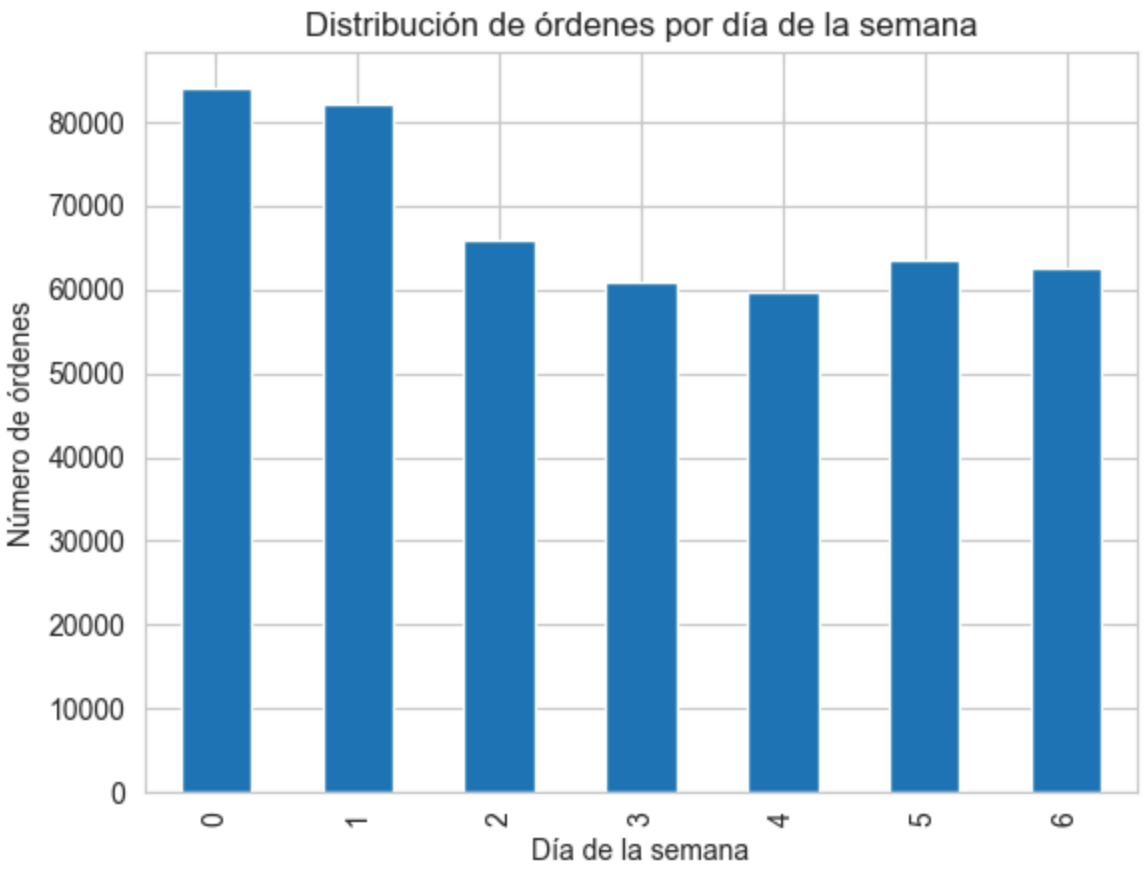
- ORDER_ID (FK)
- PRODUCT_ID (FK)
- ADD_TO_CART_ORDER
- REORDERED

2. **fct_instacart_orders**
    - ORDER_ID (PK)
    - USER_ID (PK)
    - ORDER_NUMBER
    - ORDER_DOW
    - ORDER_HOUR_OF_DAY
    - DAYS_SINCE_PRIOR_ORDER

Preguntas

1. Comportamiento de compra según día de la semana

- El gráfico muestra que el día con más órdenes es el **domingo (0)**, seguido del **sábado (6)**. Los días de la semana (lunes a viernes) tienen menos órdenes en comparación con los fines de semana.
- Los **plátanos (Banana)** son los más comprados en casi todos los días, seguidos de productos como **fresas orgánicas (Organic Strawberries)**, **espinacas orgánicas (Organic Baby Spinach)**, y **aguacates orgánicos (Organic Hass Avocado)**.

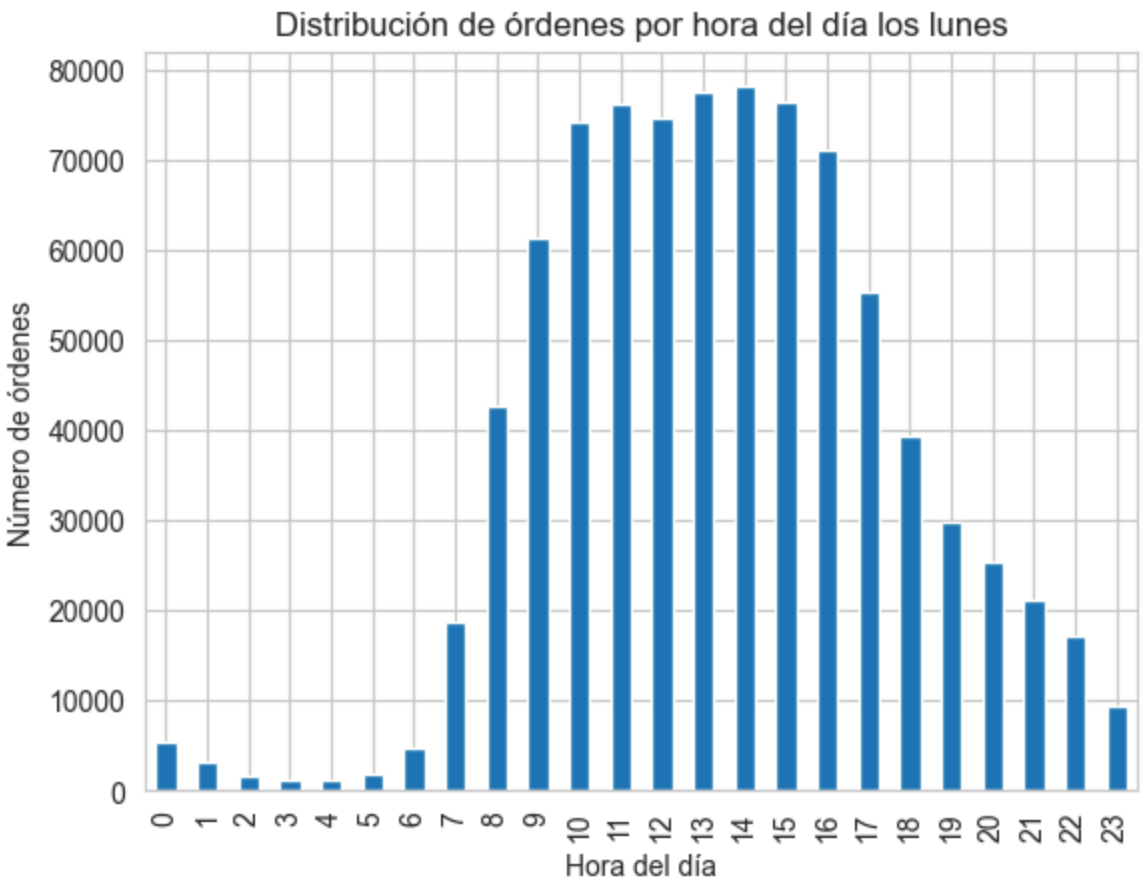


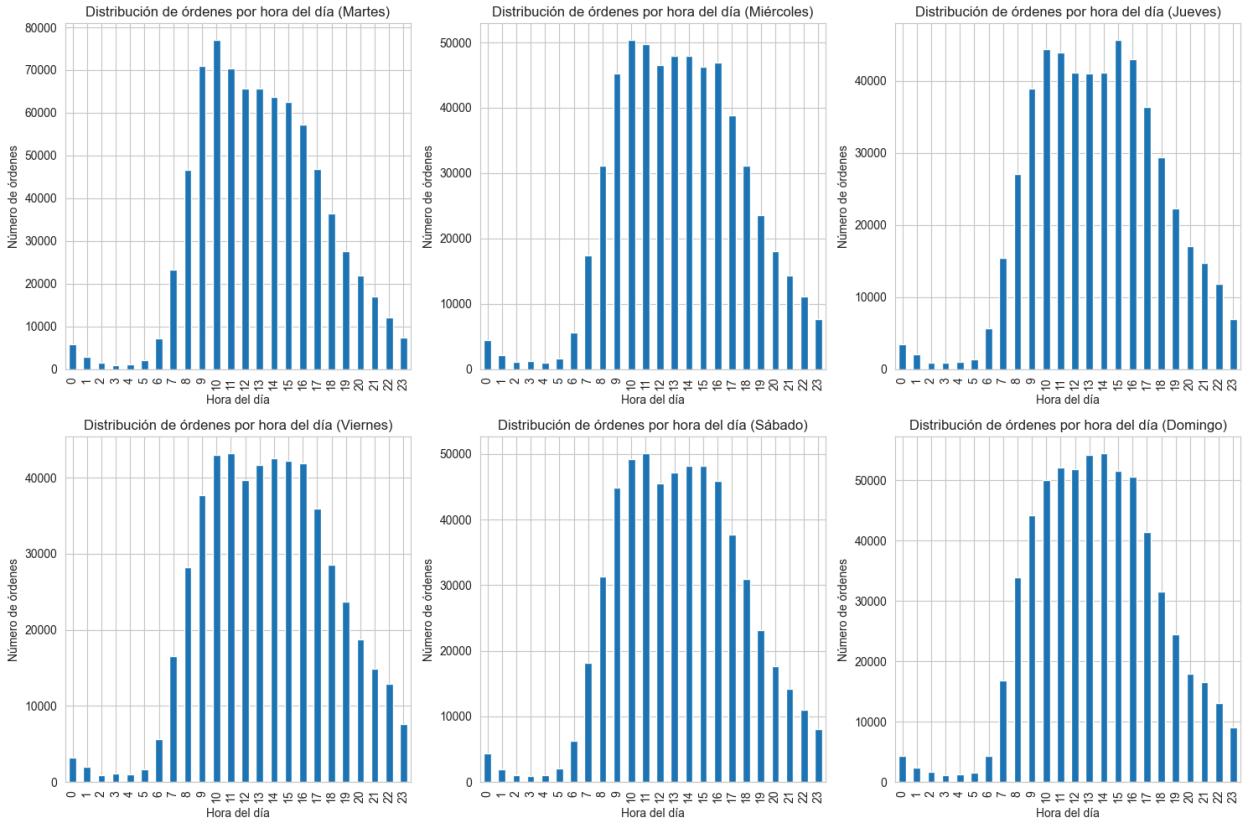
2. Comportamiento de compra según hora del día

- El gráfico muestra que la mayoría de las órdenes se realizan entre las **8:00 y las 18:00 horas**, con un pico alrededor de las **10:00 a.m.**. Las horas con menos órdenes son las primeras horas de la mañana (antes de las 8:00) y las últimas horas de la noche (después de las 20:00).
- Se analizan los productos más comprados en diferentes franjas horarias (mañana, tarde, y de 9 a 17 horas). Los productos más populares en todas las franjas horarias son los mismos: **plátanos**, **fresas orgánicas**, **espinacas orgánicas**, y **aguacates orgánicos**.

3. Comportamiento según hora del día y día de la semana

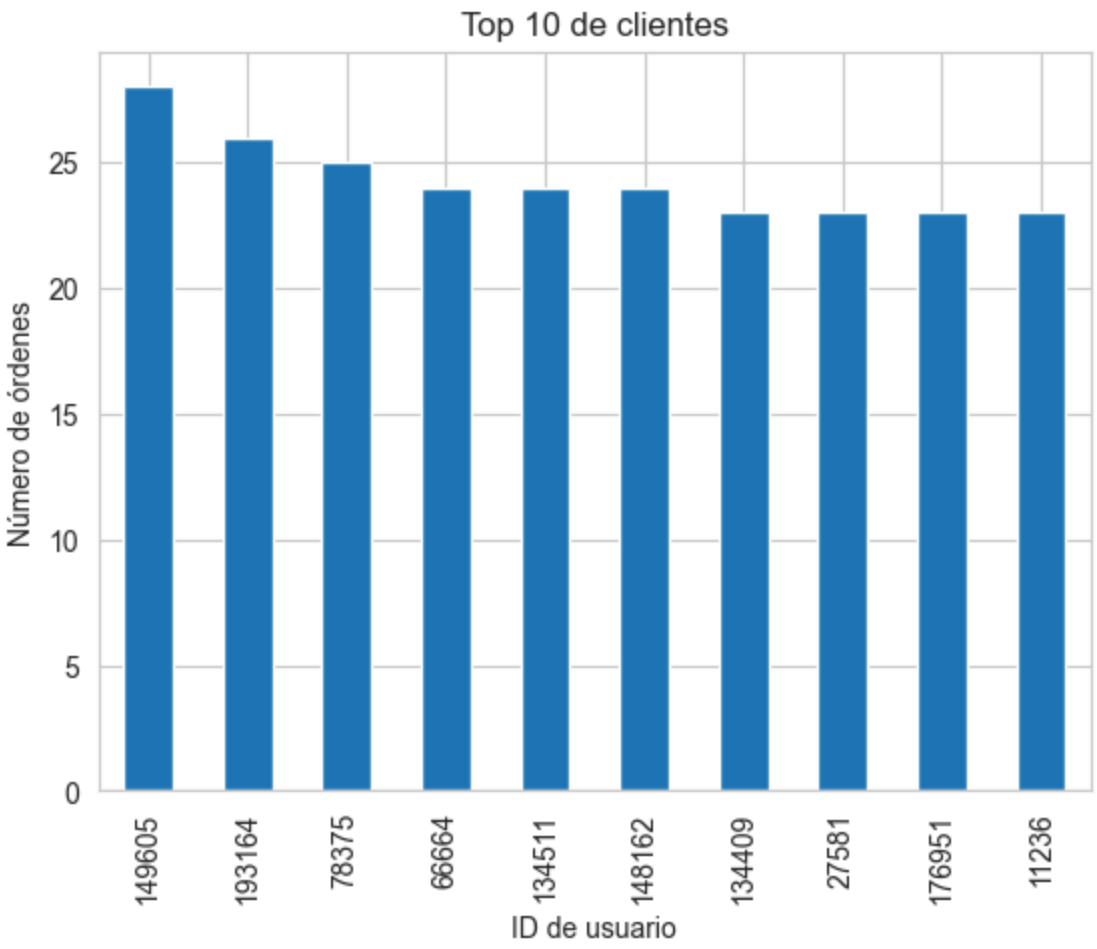
- Los patrones de compra por hora son similares en todos los días de la semana, con un pico en las horas de la mañana (alrededor de las 10:00 a.m.). Sin embargo, los fines de semana (sábado y domingo) tienen un volumen de órdenes ligeramente más alto en comparación con los días de la semana.


Distribución de órdenes por hora del día los lunes
Número de órdenes
Hora del día
0
10000
20000
30000
40000
50000
60000
70000
80000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23



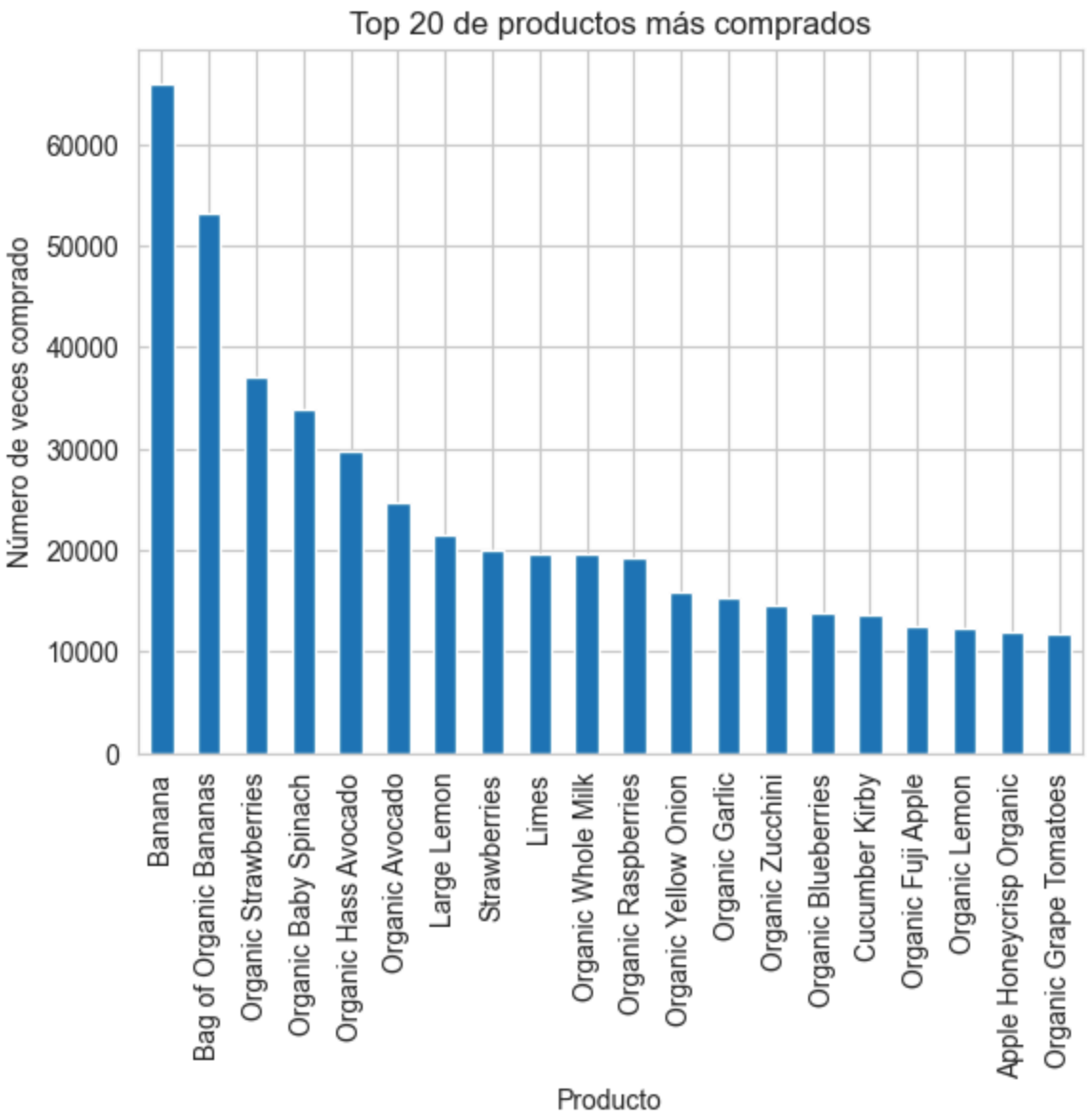
4. Distribución de las órdenes hechas por los clientes

- En promedio, un cliente realiza **3.04 órdenes**. Además, del **top 10 de clientes** con más órdenes y otro del **top 10 de clientes con menos órdenes**. Esto sugiere que hay una variabilidad significativa en el número de órdenes por cliente, con algunos clientes realizando muchas más órdenes que otros.



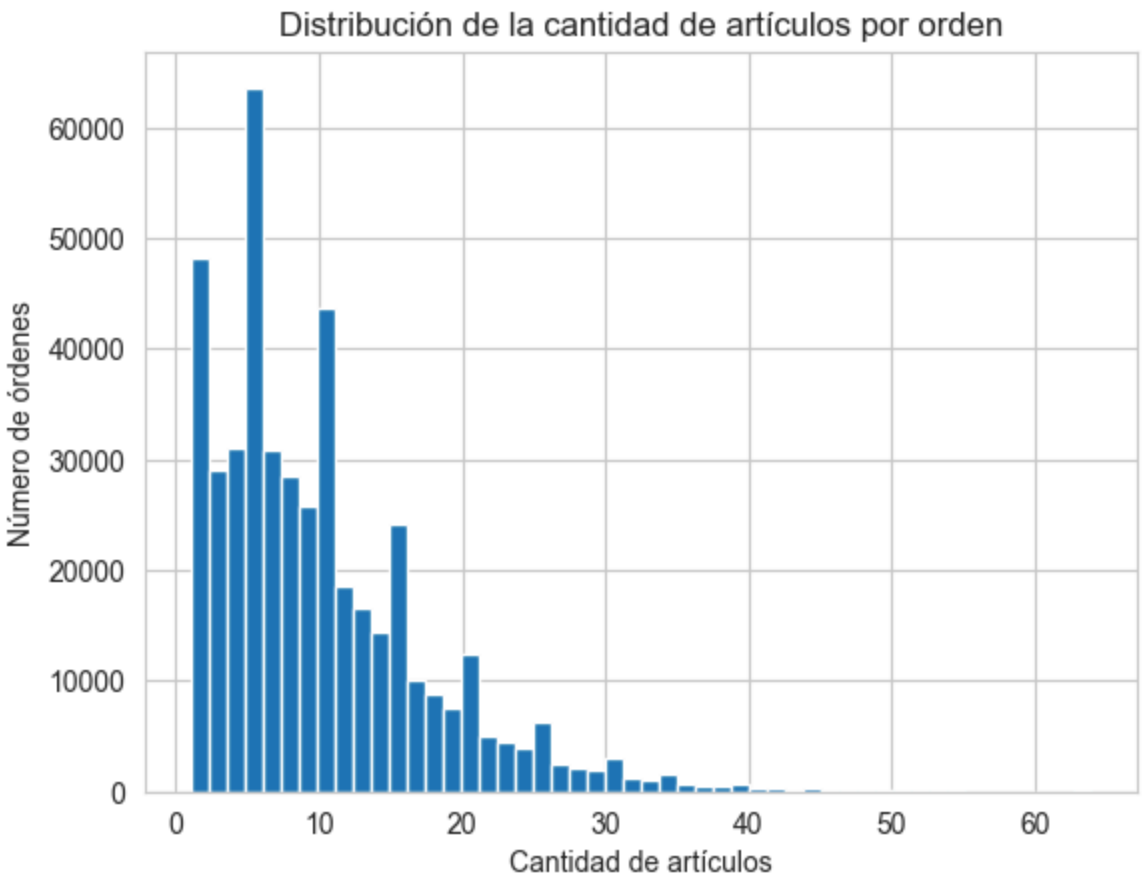
5. Top 20 productos más frecuentes

Los productos más populares son:



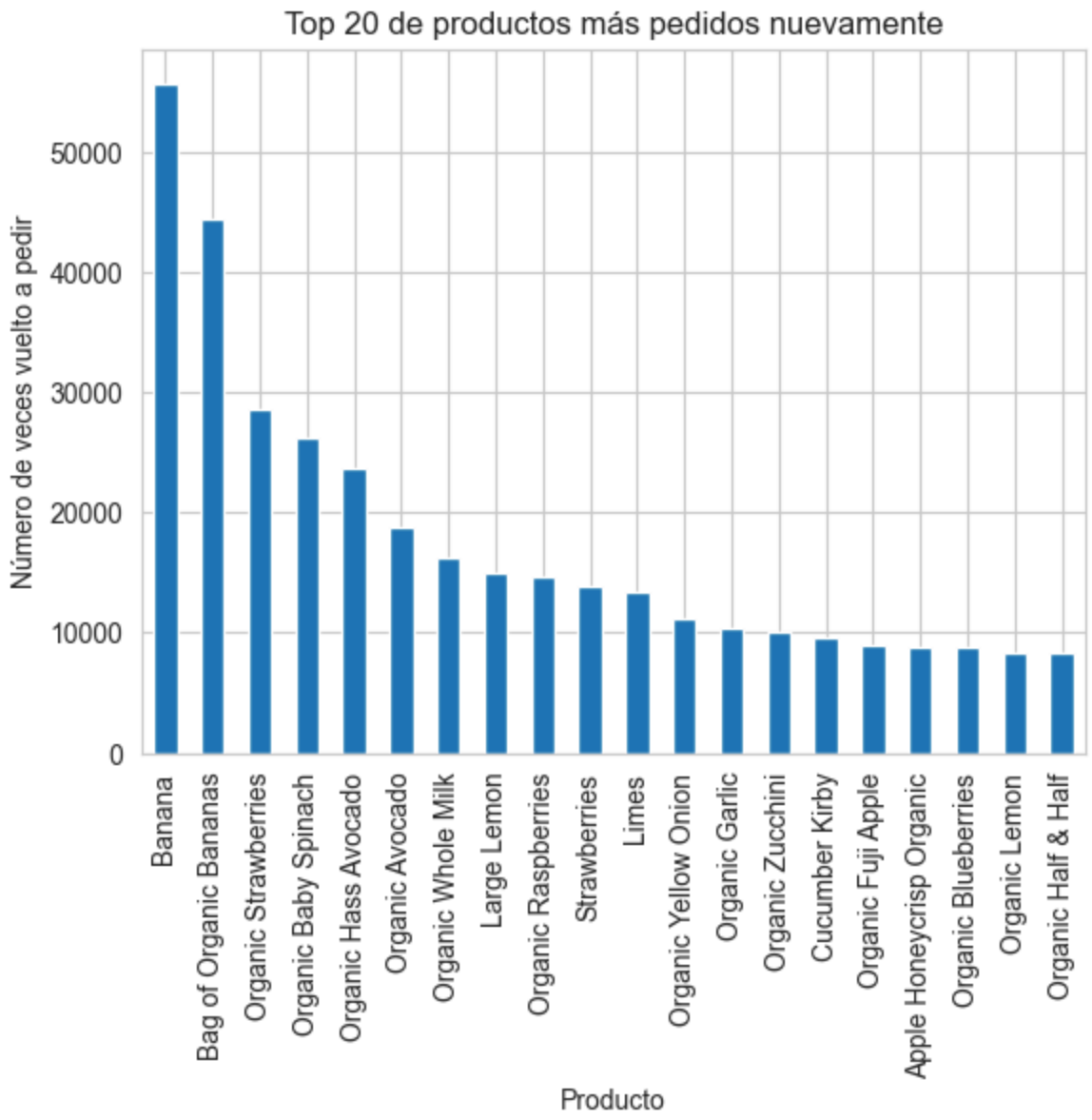
6. ¿Cuántos artículos se compran generalmente en un pedido?

- En promedio, se compran **10.1 artículos por orden**. La distribución de la cantidad de artículos por orden se muestra en un histograma, donde la mayoría de las órdenes tienen entre **5 y 15 artículos**.

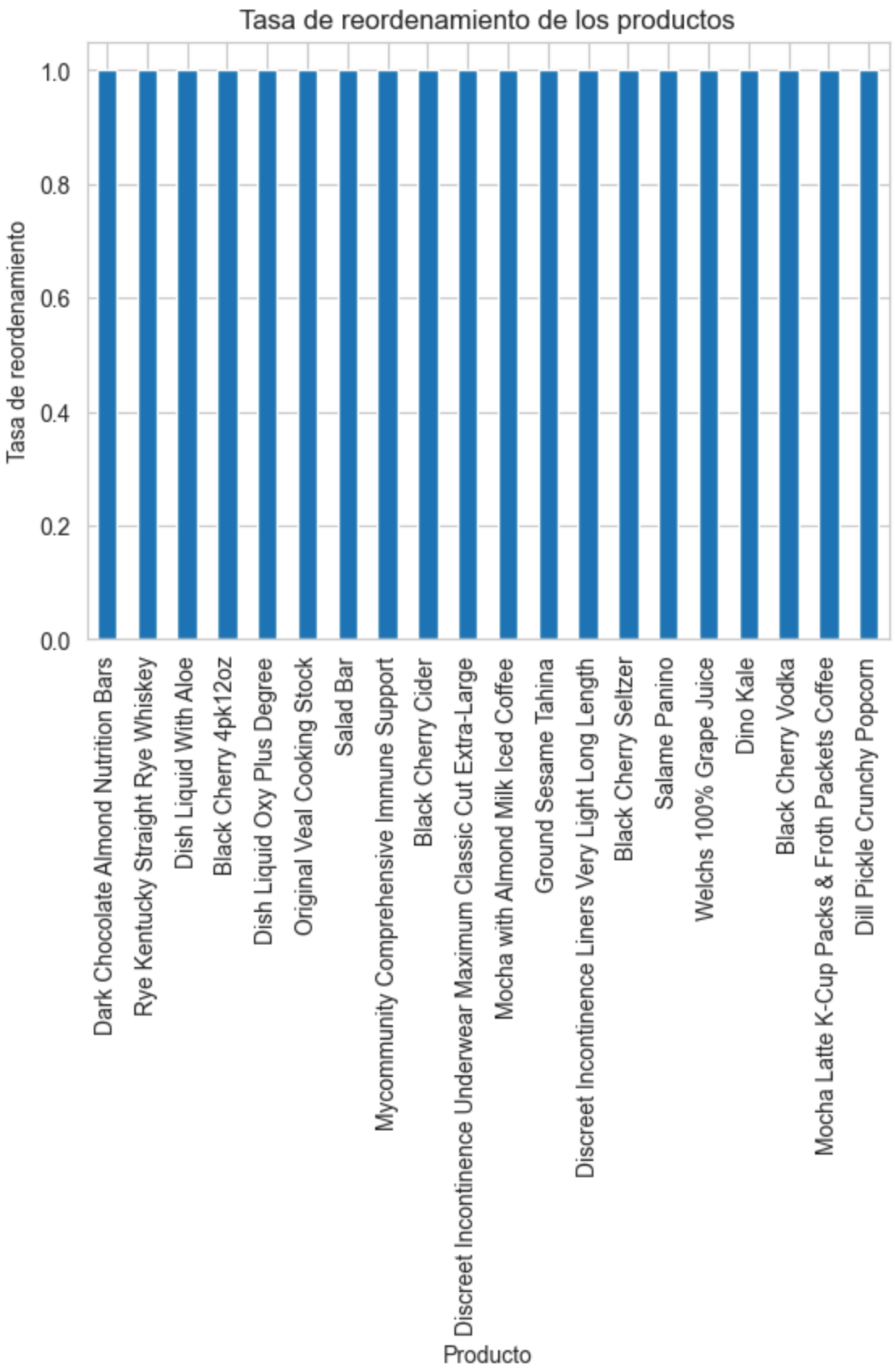


7. Top 20 artículos que se vuelven a pedir con más frecuencia

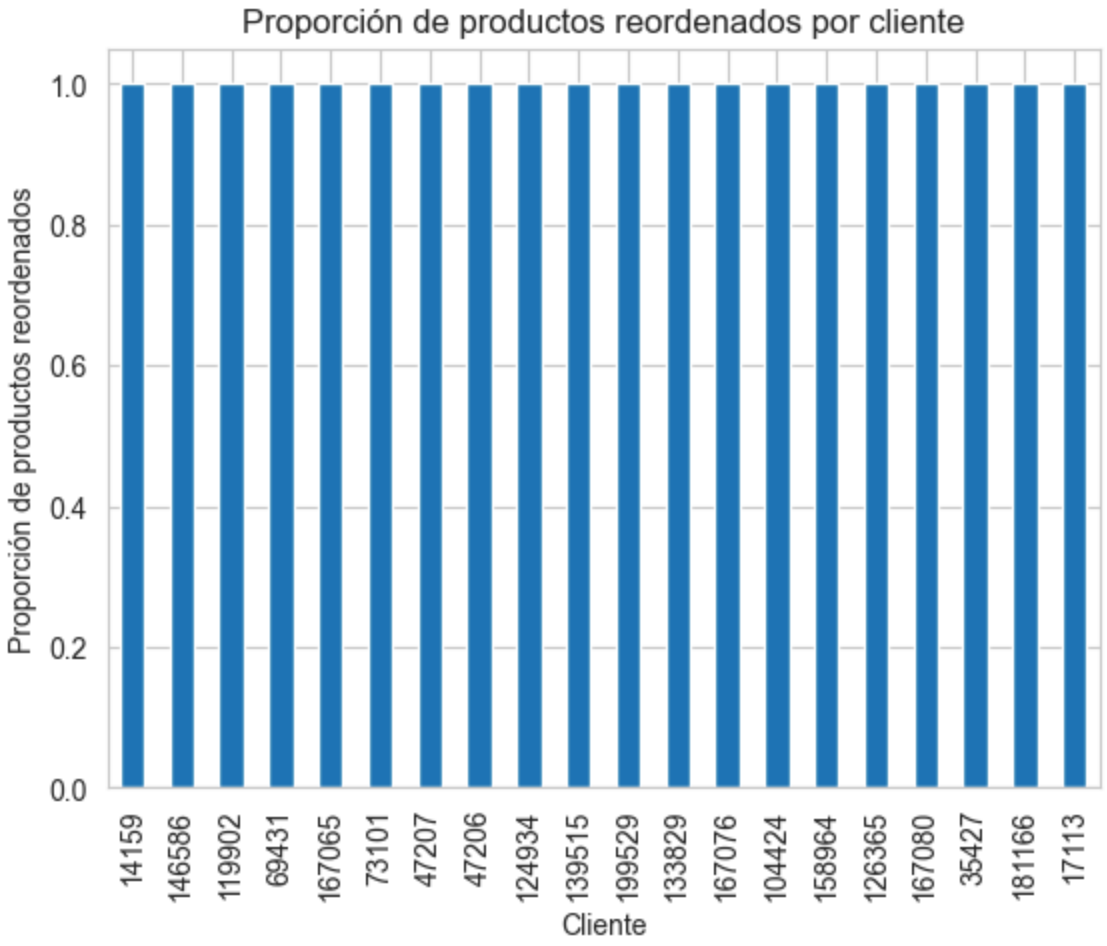
- **Productos más reordenados**:



8. Proporción de pedidos que se vuelven a pedir para cada producto

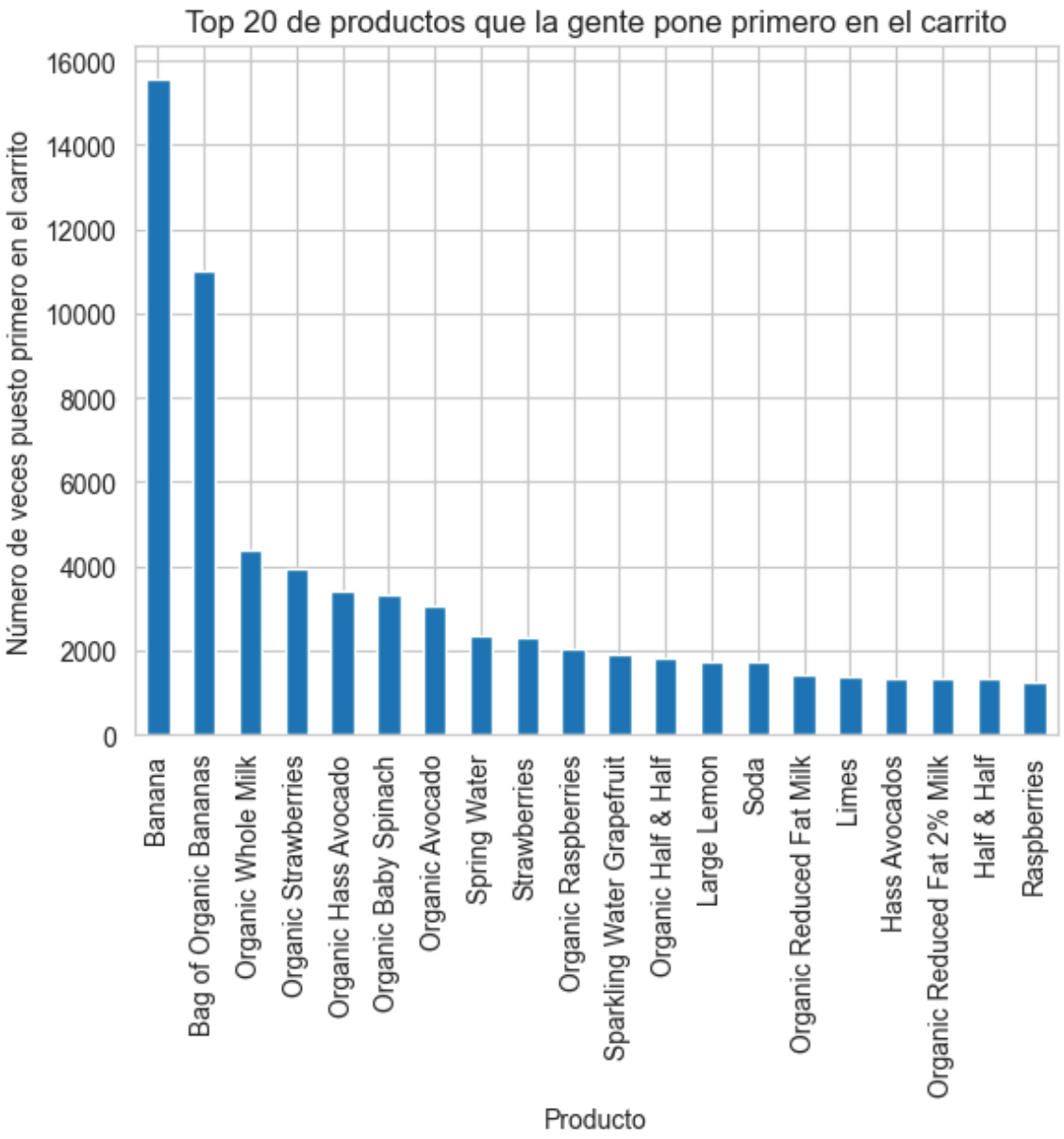

Tasa de reordenamiento de los productos
Tasa de reordenamiento
1.0
0.8
0.6
0.4
0.2
0.0
Dark Chocolate Almond Nutrition Bars
Rye Kentucky Straight Rye Whiskey
Dish Liquid With Aloe
Black Cherry 4pk12oz
Dish Liquid Oxy Plus Degree
Original Veal Cooking Stock
Salad Bar
Mycommunity Comprehensive Immune Support
Black Cherry Cider
Discreet Incontinence Underwear Maximum Classic Cut Extra-Large
Mocha with Almond Milk Iced Coffee
Ground Sesame Tahina
Discreet Incontinence Liners Very Light Long Length
Black Cherry Seltzer
Salame Panino
Welchs 100% Grape Juice
Dino Kale
Black Cherry Vodka
Mocha Latte K-Cup Packs & Froth Packets Coffee
Dill Pickle Crunchy Popcorn
Producto

9. Proporción de productos pedidos que se vuelven a pedir para cada cliente



10. Top 20 artículos que la gente pone primero en el carrito

- **Productos más comunes en primera posición**:



Conclusiones

Los fines de semana son los días con mayor actividad de compras, y los productos frescos y orgánicos son los más populares.
La mayoría de las compras se realizan en horas de la mañana, con un pico alrededor de las 10:00 a.m., y los clientes tienden a comprar en promedio 10 artículos por orden.
Los productos más reordenados son consistentes con los productos más

comprados, lo que sugiere una alta fidelidad hacia ciertos artículos.
Los clientes muestran un comportamiento heterogéneo en cuanto al número de órdenes realizadas, con un promedio de 3.04 órdenes por cliente, pero con algunos clientes realizando significativamente más compras.

Repositorio Github

https://github.com/AndresH1234/DataMining/tree/main/INSTACART