Nombre: Celic Gabriel Hernández Archundia.	**Matrícula:** 2877240
Nombre del curso: Ingeniería de datos masivos	**Nombre del profesor:** Miguel de Jesús Martínez Felipe
Módulo I	**Actividad:** Evidencia 1
Fecha: 23/09/2023	

Bibliografía:

Kaggle (s.f.). Instacart Market Basket Analysis. Recuperado el 23 de septiembre del 2023 de: https://www.kaggle.com/c/instacart-market-basket-analysis/data

Stanley, J. (2017). 3 Million Instacart Orders, Open Sourced. Recuperado el 23 de septiembre del 2023 de: https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2

dinero24. (s/f). "¿Qué es Instacart?". Recuperado el 24 de septiembre del 2023 de: https://www.dinero24.com/que-es-instacart

Simo, F. (s/f). "Instacart Mission, Vision & Values". Recuperado el 24 de septiembre del 2023 de: https://www.comparably.com/companies/instacart/mission

Instacart. (s/f). "Our Story". Recuperado el 24 de septiembre del 2023 de: https://www.instacart.com/company/about-us

THE DATA PRIVACY GROUP. (2022). "¿Qué es un diccionario de datos?". Recuperado el 24 de septiembre del 2023 de: https://thedataprivacygroup.com/es/blog/what-is-a-data-dictionary/

[Carlos Rodríguez]. Rodríguez, C. (Abril 12, 2020). "Modelos y Sistemas – Diagrama de Flujo de Datos (DFD)". Recuperado el 25 de septiembre del 2023 de: https://www.youtube.com/watch?v=vriAJOgudag&ab_channel=CarlosRodriguez

Lucidchart. (s/f). "Cómo crear un diagrama de flujo de datos". Recuperado el 25 de septiembre del 2023 de: https://www.lucidchart.com/pages/es/diagrama-de-flujo-de-datos

Instrucciones

Lee con atención plena el siguiente caso:

Instacart, es una empresa dedicada al servicio de compra y entrega en línea de comestibles, para ello cuenta con una aplicación para la realización de los pedidos. Su objetivo es facilitar el proceso de compra de despensa de sus clientes, cubriendo la demanda desde alimentos básicos, hasta los favoritos de ocasiones especiales.

La empresa ha puesto a disposición de cualquier persona un set de datos públicos de sus ventas con la finalidad de promover, entre los entusiastas de la programación y la ciencia de datos, la publicación de hallazgos que permitan mejorar su operación y habilitar una mejor experiencia de compra a sus clientes.

Elizabeth, quien recientemente ha terminado sus estudios universitarios, trabaja en una startup dedicada al sector del estudio de los datos. Junto con otras compañeras de la empresa, han decidido trabajar con los datos de Instacart en un proyecto, que tiene como objetivo principal hacer una segmentación de sus clientes e identificar los productos más vendidos en un segmento determinado. Para ello, deben seleccionar uno de los segmentos de clientes y hacer un análisis de canasta de mercado para determinar qué productos se compran frecuentemente en conjunto en un determinado segmento. Como resultado, esperan hacer una presentación con los hallazgos y una propuesta de cómo usar la información para que la empresa mejore su operación.

La propuesta del proyecto que van a desarrollar incluirá el objetivo, entendimiento del negocio, descripción de los datos disponibles y los hallazgos de un análisis exploratorio de los datos.
Imagina que tomarás el rol de Elizabeth y realiza cada una de las indicaciones, toma en cuenta que tu actividad la entregarás en tres partes diferentes (avances):

Instacart

Instacart es una empresa de entrega de comestibles de tiendas minoristas afiliadas a este. Las personas solo tienen que ingresar a la aplicación o página web de la empresa, seleccionar sus pedidos y pagar los productos. Instacart fue fundada en 2012 en la Bay Area de San Francisco por la ex empleada de Amazon Apoorva Metha, y otros miembros como Max Mullen y Brandon Leonardo, con el objetivo de ayudar a los consumidores a conseguir alimentos básicos y frescos desde la comodidad del hogar o la oficina.

Misión y Visión de Instacart.

La misión de esta empresa es la de crear un mundo donde todas las personas puedas acceder a la comida que aman y tener más tiempo para disfrutarla entre todos. Entre sus valores se encuentran los siguientes:

- Llegar más lejos juntos
- Crear nuevos mercados y oportunidades
- Preparación para el trabajo duro
- Servicio generoso

Funciones e Importancia

Instacart busca que las personas aprovechen su tiempo mientras la compra de alimentos se lleva a cabo. En vez de que las personas mismas vayan y compren su alimento, labor que puede durar bastante, Instacart permite hacer estos pedidos con ayuda de sus plataformas y servicio asociados. Siendo posible no solo programar los horarios de compra del usuario, sino que también permitiendo la opción de recogida de los productos en el establecimiento. Es decir, la persona puede escoger y comprar sus comestibles en su tienda de confianza desde la aplicación o la plataforma web y, posteriormente, recogerlos en la tienda según la hora seleccionada (disponible para ciertas ubicaciones).

La importancia de este servicio se ha visto presente en varios ámbitos, en los que nos gustaría destacar los siguientes:

- **Entrega de comestibles a domicilio:** Aunque este hecho sea un poco repetitivo, el que los compradores de Instacart recojan los productos en las tiendas asociadas o lleguen a domicilio por medio de una aplicación es un gran servicio que permite a las personas tener más tiempo para ellas mismas o para su trabajo.

- **Amplia red de tiendas asociadas:** Instacart ha establecido asociaciones con una amplia variedad de supermercados y minoristas, lo que le permite ofrecer una amplia selección de productos a sus clientes. Entre los socios de Instacart se encuentran cadenas de supermercados como Costco, Safeway, Kroger, Walmart y muchos otros.

- **Modelo de negocio de gig economy:** Instacart opera en gran medida mediante un modelo de negocio de gig economy, contratando a "Shoppers" que realizan las compras y entregas a los clientes. Esto les permite trabajar de manera flexible y ganar dinero como trabajadores independientes.

- **Expansión y crecimiento:** Instacart ha experimentado un rápido crecimiento desde su fundación. Ha ampliado sus operaciones a lo largo de los años y ahora opera en miles de ciudades en los Estados Unidos y Canadá. También ha introducido servicios como Instacart Express, que ofrece entregas ilimitadas por una tarifa mensual, y ha expandido su oferta de productos más allá de los comestibles para incluir productos de farmacia y minoristas de electrónica, entre otros.

- **Importancia durante la pandemia de COVID-19:** Durante la pandemia de COVID-19, Instacart se convirtió en un servicio esencial para muchas personas que buscaban evitar las tiendas físicas y minimizar el contacto social. La demanda de entregas de comestibles a domicilio aumentó significativamente, lo que llevó a un mayor crecimiento y expansión de Instacart.

Diccionarios de Datos

Los diccionarios de datos son metadatos que brindan información sobre los elementos que lo componen. Básicamente, son datos dentro de un repositorio, tabla, o almacén que describen elementos relevantes de las bases de datos con las que los usuarios trabajan.

Según The Data Privacy Group, “un diccionario de datos bien mantenido es una herramienta fundamental para garantizar datos coherentes y precisos en toda una organización, permitiendo a los usuarios comprender el significado y la finalidad de los datos”. Entre las principales ventajas que estos ofrecen se encuentran:

- Importantes puntos de referencia para cualquier persona que acceda a los datos y los analice.
- Garantiza la coherencia de los datos de la organización.
- Reduce el riesgo de errores referentes a la interpretación de datos.
- Proporciona una estructura organizada y una visión general de todos los elementos de la base de datos.
- Ayuda a garantizar el cumplimiento de cualquier norma y reglamento de calidad de datos existente.

A continuación, se presenta el diccionario de datos de cada archivo proporcionado por Instacart:

Archivo 1: “aisles”

Aisle					
Nombre del campo	Descripción	Tipo de dato	Longitud	Índice	Tipo de Índice
aisle_id	Identificador de pasillos o secciones de comestibles	INT	No aplica	Sí	Primario
aisle	Nombre de la sección de comestibles	VARCHAR	50	Sí	Único

Archivo 2: “departments”

Departments					
Nombre del campo	Descripción	Tipo de dato	Longitud	Índice	Tipo de Índice
department_id	Identificador de departamentos o categorías de productos	INT	No aplica	Sí	Primario
department	Nombre del departamento	VARCHAR	30	Sí	Único

Archivo 3: “order_products__prior”

Order_products__prior					
Nombre del campo	Descripción	Tipo de dato	Longitud	Índice	Tipo de Índice
order_id	Identificador de orden	INT	No aplica	Sí	Compuesto
product_id	Identificador de producto	INT	No aplica	Sí	Primario
add_to_cart_order	Número que identifica el orden en que se añadieron	INT	No aplica	Sí	Compuesto

	los productos al carrito				
reordered	El cliente reordenó el producto (0 o 1)	INT	No aplica	Sí	Compuesto

Archivo 4: "order_products__train"

Order_products__train					
Nombre del campo	Descripción	Tipo de dato	Longitud	Índice	Tipo de Índice
Order_id	Identificador de orden respecto al conjunto de productos	INT	No aplica	Sí	Compuesto
Product_id	Identificador de producto	INT	No aplica	Sí	Primario
Add_to_cart_order	Número que identifica el orden en que se añadieron los productos al carrito	INT	No aplica	Sí	Compuesto
reordered	El cliente reordenó el producto (0 o 1)	INT	No aplica	Sí	Compuesto

Archivo 5: "orders"

Orders					
Nombre del campo	Descripción	Tipo de dato	Longitud	Índice	Tipo de Índice
order_id	Identificador de orden único	INT	No aplica	Sí	Primario
user_id	Identificador de usuario	INT	No aplica	Sí	Compuesto
eval_set	Tipo de evaluación	VARCHAR	5	Sí	Compuesto

order_number	Número de orden	INT	No aplica	Sí	Compuesto
order_dow	Día de la semana de la orden	INT	No aplica	Sí	Compuesto
order_hour_of_day	Hora del día de la orden	INT	No aplica	Sí	Compuesto
days_since_prior_order	Días desde la orden anterior	FLOAT	No aplica	Sí	Compuesto

Archivo 6: "products"

Products					
Nombre del campo	Descripción	Tipo de dato	Longitud	Índice	Tipo de Índice
product_id	Identificador del producto	INT	No aplica	Sí	Secundario
product_name	Nombre del producto	VARCHAR	50	Sí	Primario
aisle_id	Identificador de pasillo o sección de comestibles	INT	No aplica	Sí	Secundario
department_id	Identificador del departamento	INT	No aplica	Sí	Secundario

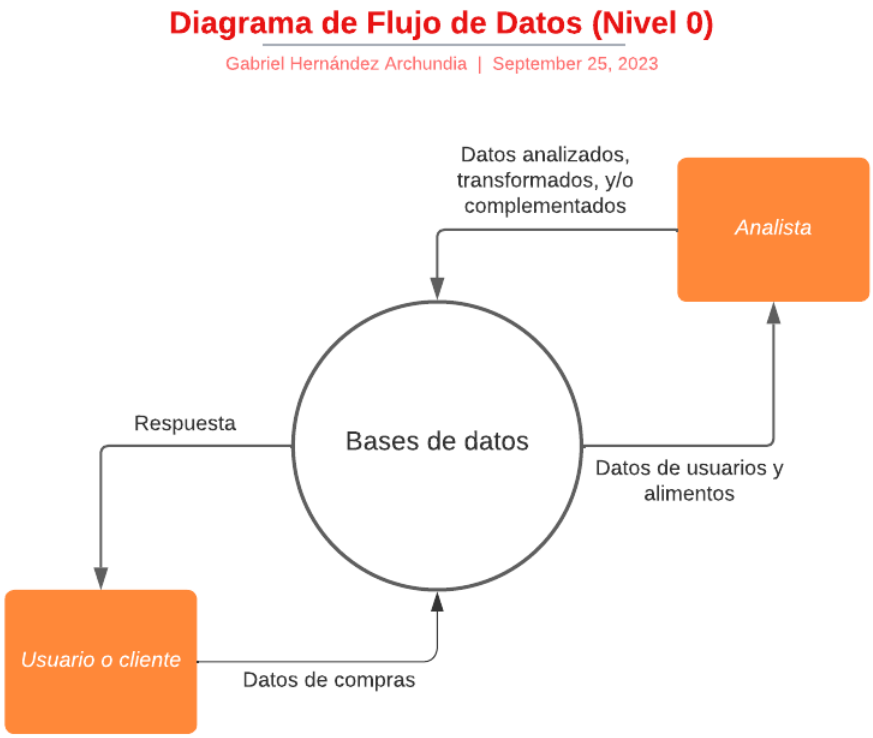
Archivo 7: "sample_submission"

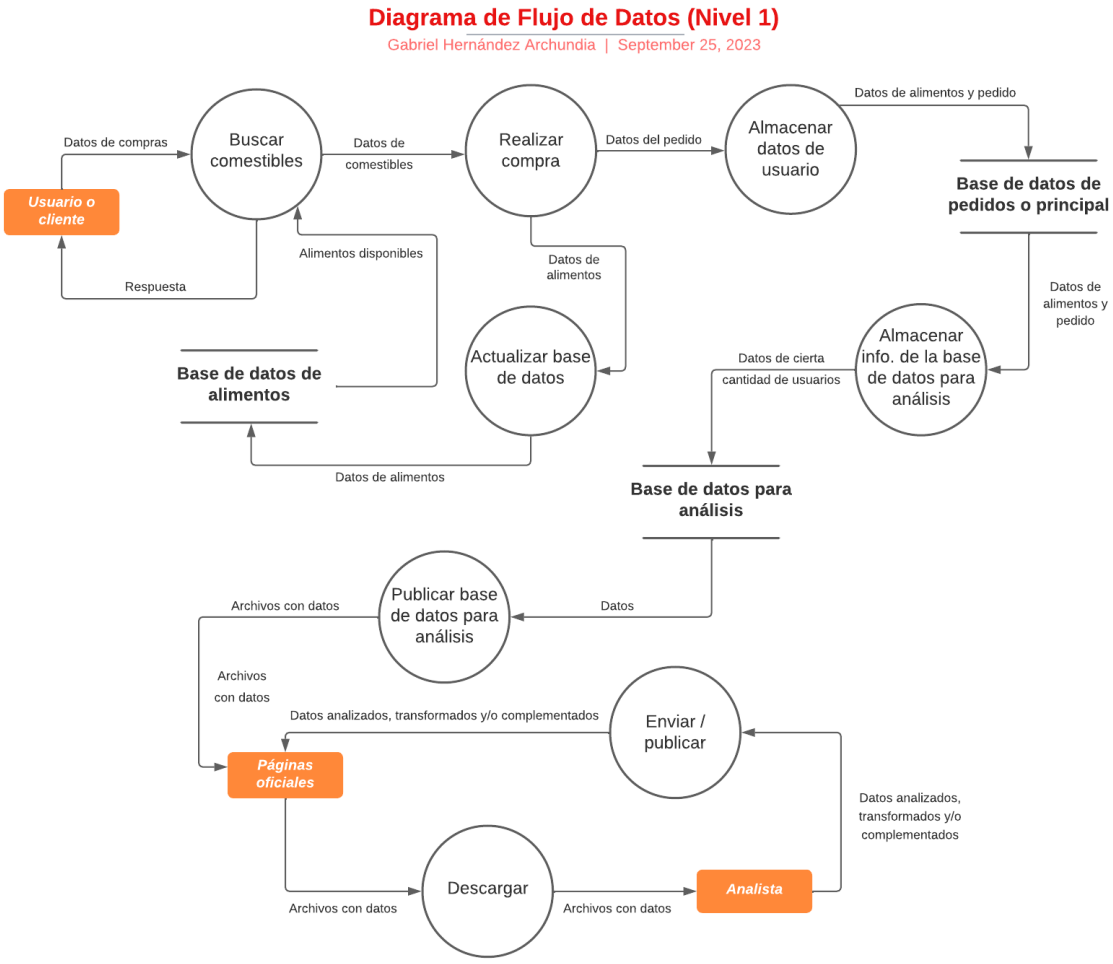
Sample_Submission					
Nombre del campo	Descripción	Tipo de dato	Longitud	Índice	Tipo de Índice
order_id	Identificador de la orden	INT	No aplica	Sí	Primario
Products	Número de productos	INT	No aplica	Sí	Ordinario

Diagramas de flujo de datos

Un diagrama de flujo de datos es una representación visual de cómo fluyen los datos a través de un sistema; esto permite saber de dónde vienen los datos, hacia dónde se dirigen y cómo o dónde se almacenan. Esta representación visual se logra con la ayuda de ciertos símbolos, los cuáles pueden indicar: procesos, como una actividad de negocios en la que ocurre una manipulación de datos o su transformación; flujo de datos, representados con flechas; entidades externas, que pueden ser personas, sistemas o una aplicación; y almacenes de datos, que muestran dónde se almacenan o se producen los datos con relación al proceso.

Los diagramas de datos son creados originalmente a partir de un modelo general (nivel 0) que posteriormente se detalla en un modelo más desarrollado (nivel 1). Estos serían los diagramas de flujo de datos para la empresa Instacart:





Los archivos proporcionados por Instacart nos revelan cierta información que es de utilidad para el entendimiento del cliente y encontrar patrones que revelen nuevas tendencias o sucesos actuales. A continuación, se presenta la información relevante contenida en cada archivo:

Análisis archivo 1: En el primer archivo podemos ver el contenido de las secciones de comestibles o pasillos (“aisle”) junto con su correspondiente identificador de sección (“aisle_id”). Algunos ejemplos de secciones de comestibles encontrados dentro de este archivo son los siguientes:

- “Prepared soups salads”
- “Specially cheeses”
- “Energy granola bars”
- “Instan soups”
- “Marinades meat preparation”

La utilidad de estos datos se verá presente en el trabajo conjunto de las bases de datos; cuando juntemos esta información junto con otras para encontrar patrones de compras de los clientes, por ejemplo.

Análisis archivo 2: El segundo archivo contiene la información de los distintos departamentos de la tienda en línea de Instacart ("department"); las cuales son el hogar de las secciones de comida. Cada departamento cuenta con su correspondiente identificador ("department_id") y cada departamento es único dentro de esta base de datos. Algunos departamentos que podemos encontrar son los siguientes:

- "Frozen"
- "Bakery"
- "International"
- "Alcohol"
- "Breakfast"
- "Snacks"
- "Personal care"

Al igual que con el primer archivo, esta base de datos tomaría un rol más importante cuando se use en conjunto con las demás, ya se para búsqueda de información, búsqueda de patrones, etc.

Análisis archivo 3: El tercer archivo contiene un poco más de información que sus antecesores. En él podemos encontrar 4 columnas, las cuales describen el identificador de orden ("order_id"), el identificador del producto ("product_id"), el orden en que los productos se añadieron al carrito ("add_to_cart_order"), y si hubo algún reorden o no ("redordered"), representada con 0 y 1; seguramente el número cero siendo "falso" y el número uno como "verdadero".

La información de Instacart va cobrando más sentido conforme analizamos los archivos proporcionados. La principal utilidad que se puede obtener de esta base de datos son los identificadores de orden de los usuarios y los de los productos. Sin embargo, hay que complementar esta información para poder crear algo más interesante. Cabe aclarar que esta información es respecto a los productos de tipo "prior". Esta información será descrita de mejor forma en los próximos archivos.

Análisis archivo 4: El cuarto archivo es muy similar a la vista en el tercer archivo. Contiene exactamente las mismas columnas vistas con anterioridad; identificador de orden ("order_id"), identificador de producto ("product_id"), orden en que los productos se añadieron al carrito ("add_to_cart_order"), y un indicador de reorden ("reordered"). Los productos de esta base de datos son los de tipo "train". Esta información será descrita de mejor forma en los próximos archivos.

Análisis archivo 5: El archivo 5 me parece el más interesante de todos debido a su contenido. En este se pueden encontrar aún más columnas, las cuales son:

- Identificador de orden (“order_id”).
- Identificador de usuario (“user_id”).
- Tipo de evaluación (“eval_set”).
- Número de orden (“order_number”).
- Día de la semana de la orden (“order_dow”).
- Hora del día de la orden (“order_hour_of_day”).
- Días desde la orden anterior (“days_since_prior_order”).

La mayoría despliega alguna respuesta de tipo entero (INT). A continuación, se describiría un poco más a detalle el contenido de cada columna vista en esta base de datos:

- “order_id”: Es simplemente el identificador de orden en que se realizó cierta compra por el usuario. Este identificador se representa con números.

- “user_id”: Dentro de esta columna se pueden observar cuántas órdenes emitió el usuario a lo largo de las filas. No son datos únicos, ya que se repiten en varias ocasiones.

- “eval_set”: Esta columna (por lo visto) solo puede contener tres tipos de “VARCHAR”: “prior”, “train”, o “test”:

    - *prior:* Esta etiqueta se usa para los registros que corresponden a pedidos anteriores (prior orders) de los clientes. En otras palabras, son pedidos históricos que los clientes han realizado antes de un punto de referencia específico. Estos registros se utilizan comúnmente para analizar el historial de compras de los clientes y comprender sus hábitos de compra pasados.

    - *train:* Esta etiqueta se usa para los registros que forman parte del conjunto de entrenamiento (train set). El conjunto de entrenamiento suele utilizarse para desarrollar y ajustar modelos de aprendizaje automático. Contiene ejemplos históricos de compras de los clientes junto con información sobre si ciertos productos fueron comprados en un pedido posterior. Los modelos se entrenan utilizando estos datos para predecir compras futuras.

    - *test:* Esta etiqueta se utiliza para los registros que forman parte del conjunto de pruebas (test set). El conjunto de pruebas se utiliza para evaluar el rendimiento de los modelos de aprendizaje automático desarrollados en el conjunto de entrenamiento. Contiene ejemplos de compras de clientes, pero generalmente no incluye información sobre las compras posteriores. Los modelos se utilizan para predecir las compras futuras en función de estos datos de prueba.

- “order_number”: Es básicamente el número de la orden de cada usuario en cierto tiempo. Es un valor que incrementa automáticamente de 1 en 1 iniciando en 1 para cada identificador de usuario.

- “order_dow”: Día de la semana en que la orden fue emitida (posiblemente siendo cero el número correspondiente al domingo).
- “order_hour_of_day”: La hora a la que se concretó la orden del cliente. Esta es representada por un número de tipo entero (INT).

- “days_since_prior_order”: Son los días que han pasado desde la última orden del cliente. Aquí se pueden encontrar números de tipo flotante y “NaN” (Not a Number).

Análisis archivo 6: En el sexto archivo podemos encontrar detalles más específicos sobre los productos. En la base de datos podemos encontrar las siguientes columnas: Identificador del producto (“product_id”), útil para encontrar los productos de forma más rápida; nombre del producto (“producto_name”), para saber de qué producto se trata; identificador de sección o pasillo (“aisle_id”), para saber a qué pasillo pertenece dicho producto (y con quién lo comparte); y el identificador del departamento (“department_id”), que es otra forma de localizar al producto.

Esta información por separado es más que nada para lograr identificar o ubicar los productos de Instacart dentro de su base de datos. Estos datos de ser usados en conjunto con los anteriores podrían ayudar en la búsqueda de patrones o predicciones de manera personalizada. Es decir, solo sobre ciertos productos de interés.

Análisis archivo 7: El último archivo, llamado “sample_submission”, contiene un identificador de orden (“order_id”) y una columna titulada “products”. La columna "order_id" probablemente contiene identificadores únicos para cada pedido, y la columna "products" parece estar relacionada con los productos que están asociados a cada pedido. Los números que se pueden observar en la columna "products" podrían ser identificadores de productos o algún tipo de código que hace referencia a los productos específicos que fueron incluidos en cada pedido. Sin embargo, hay que tomar esta información con discreción.

Recapitulación y entendimiento del negocio

Recapitulando, Instacart es una plataforma de entrega de comestibles y productos de abarrotes a domicilio que permite a los usuarios ordenar alimentos y otros productos de tiendas locales y cadenas de supermercados a través de su aplicación o sitio web. Y su misión está en línea con la evolución del comercio minorista y la tecnología digital para hacer que las compras de comestibles sean más convenientes y accesibles para los consumidores.

Objetivos del proyecto

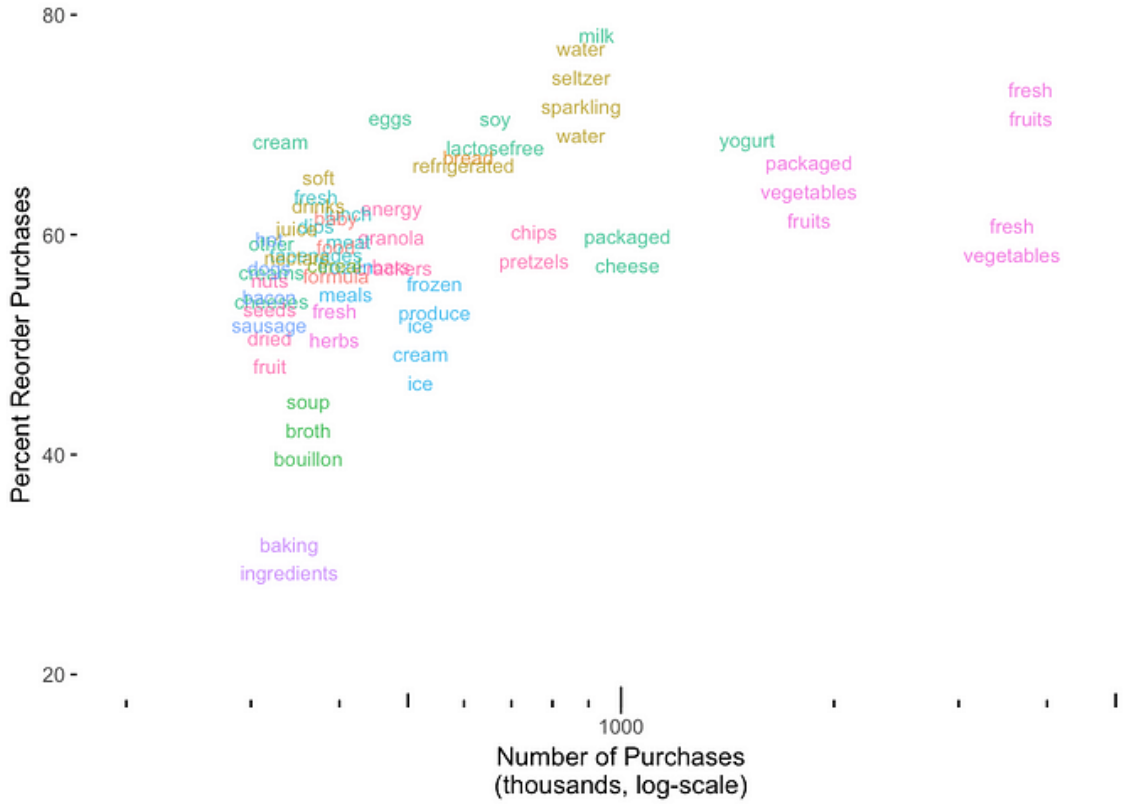
Los objetivos de este proyecto de análisis es entender el negocio de Instacart, brindar una buena descripción de los datos disponibles y hacer hallazgos derivado del análisis exploratorio de los datos.

Descripción general y utilidad

En este primer trabajo, pudimos ver los archivos de datos proporcionados por la empresa Instacart. Muchos de estos datos necesitan ser usados de manera conjunta para el análisis de su mercado. Entre los datos más relevantes para ello, podemos encontrar los siguientes:

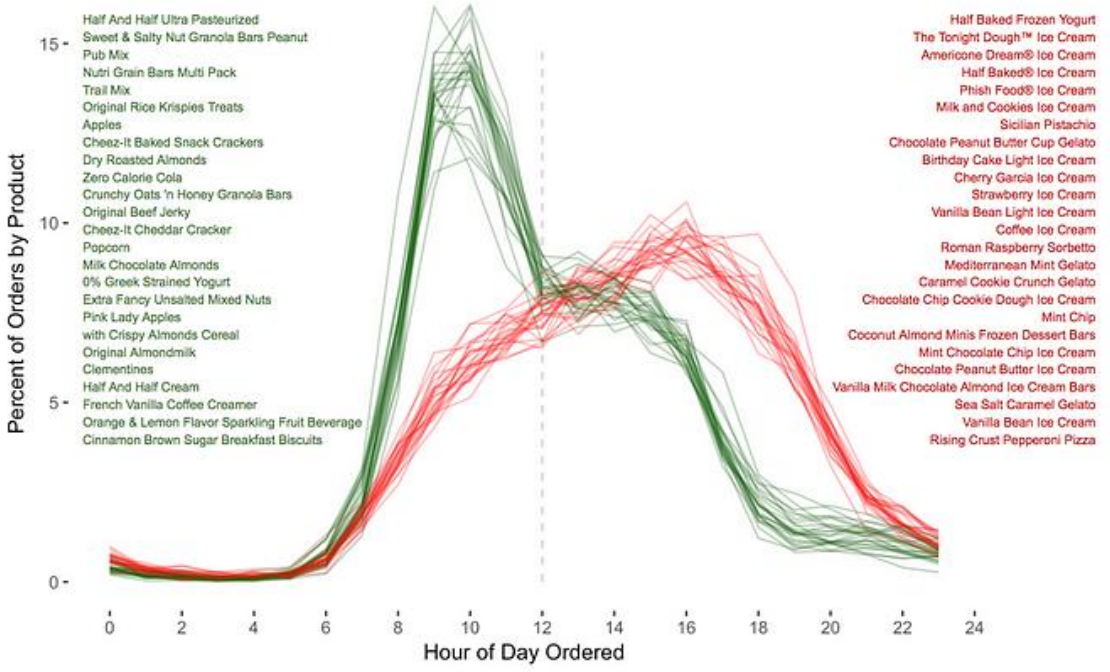
- “order_id”
- “order_number”
- “order_dow”
- “order_hour_of_day”
- “add_to_cart_order”
- “product_id”
- “product_name”
- “aisle”

Con datos como estos, se pueden encontrar patrones dentro de una base de una base de datos. Por ejemplo, la siguiente gráfica muestra que los pasillos o secciones de comestibles más comunes o concurridos son más propensos a ser reordenados.



En este caso se puede observar que las frutas son reordenadas más frecuentemente que los vegetales (tal vez porque los vegetales son más usados para hacer complementar recetas) y que las sopas e ingredientes para hornear son menos propensos a ser comprados de nuevo (tal vez porque no son tan necesarios en la alimentación del día a día).

Otra gráfica que se ha podido obtener con ayuda de los datos es la siguiente:



En el que se puede ver que los bocadillos más sanos y productos básicos se compran con frecuencia en las mañanas. Mientras que los helados (espacialmente el “Half Baked” y “The Tonight Dough”) son comúnmente solicitados por los clientes en horarios de la tarde o en la noche.

Estos son solo un par de ejemplos de la información que se puede obtener mediante el análisis de datos. El propósito principal de este proyecto es descubrir más patrones como estos y/o conseguir información relevante. Esta información se verá complementada en próximos avances.