BODY TECH®

↘ Exploración

BODYTECH: Más fuerte cada dato

+ Julio 2024

ÍNDICE

INTRODUCCIÓN

ANÁLISIS UNIVARIABLE

ANÁLISIS MULTIVARIABLE

SISTEMA EMPRESARIAL CONFUSO

DATAFRAME DE LEADS

DATAFRAME DE DEALS

DATAFRAME DE SELLS

DATAFRAME DE DEALS

DATAFRAME DE INVOICES

CONCLUSIONES

INTRODUCCIÓN

BodyTech, líder en el sector de gimnasios y bienestar, enfrenta el desafío de gestionar eficientemente sus recursos y optimizar la oferta de servicios para maximizar sus ingresos. Nuestra propuesta, "BodyTech, más fuerte cada dato", busca proporcionar una solución integral que combine datos históricos y análisis en tiempo real para mejorar la capacidad de predicción de picos de demanda y la identificación de oportunidades de venta, mejorando así la planificación y gestión de recursos de la empresa.

Nivel de Solución

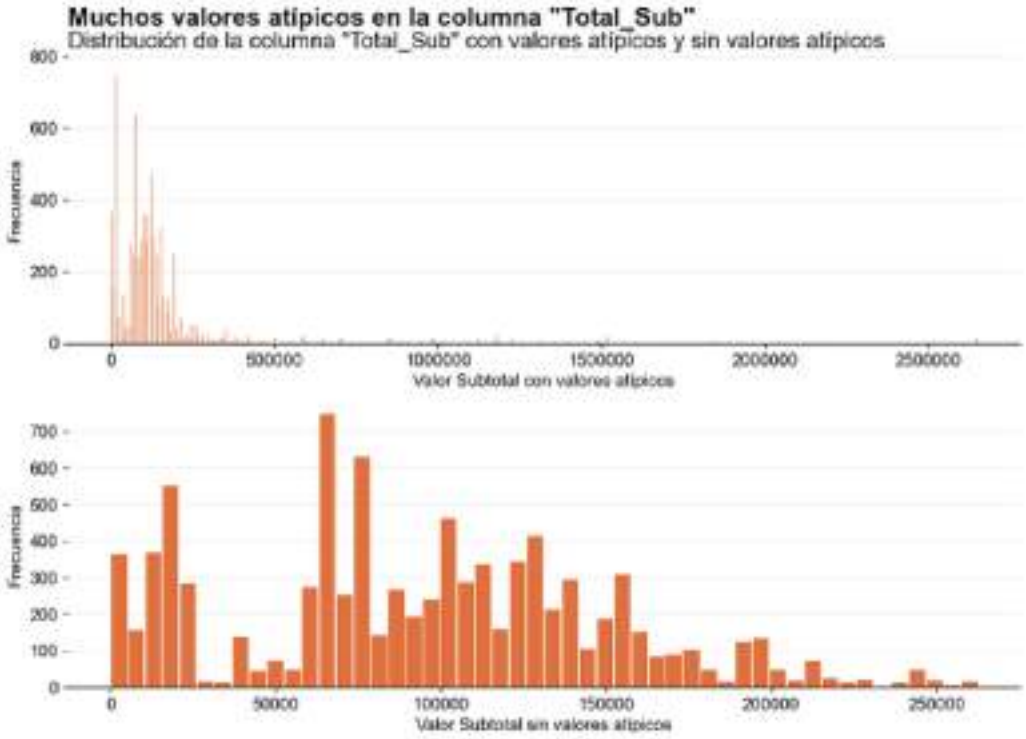
Sugeriremos ofertas comerciales personalizadas para cada pico de demanda, basándonos en el tipo de cliente (Nuevos, Renovados, Recuperados) y el tipo de servicio (Recurrentes y Único Pago). Estas ofertas estarán diseñadas para maximizar las oportunidades de venta y mejorar la satisfacción del cliente, considerando el historial de promociones y su efectividad, así como las preferencias y comportamientos de los clientes.

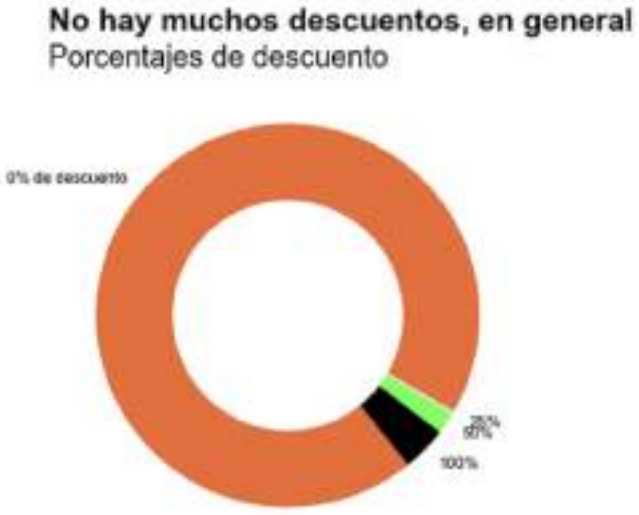
ANÁLISIS UNIVARIABLE

Comenzamos analizando la columna con los valores totales, pues encontramos registros de pagos cuyo valor era $0 COP que podía entorpecer el entrenamiento del algoritmo de Machine Learning. Estos valores correspondían al siguiente tipo de pago:

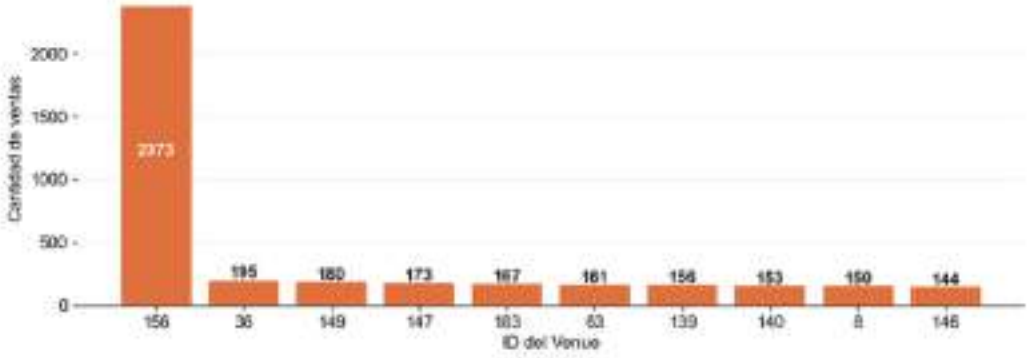
Tipo de pago	Cantidad de pagos
Membresía	211
Débito Gold Grupal x2	13
Débito Tradicional Grupal x2	11
Débito Mensual Gold - Gran Manzana	9
Débito Mensual Corporativo	5
Débito Gold Grupal x3	2
Débito Gold Grupal x4	1
Multa	1

Dado que hay tantos registros de la columna `Total_Sub` con valor 0, y que la mayoría de ellos, 211, corresponden a `Membresía`, los eliminamos para efectos del análisis.





El venue 156 vende por 13 venues
Los venues con mayor volumen de ventas



Muchos miembros desconocidos

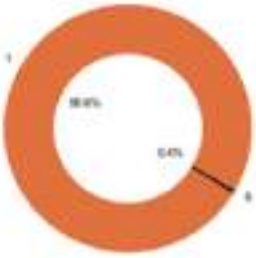
Porcentaje de estado de los miembros

Nuestros aliados

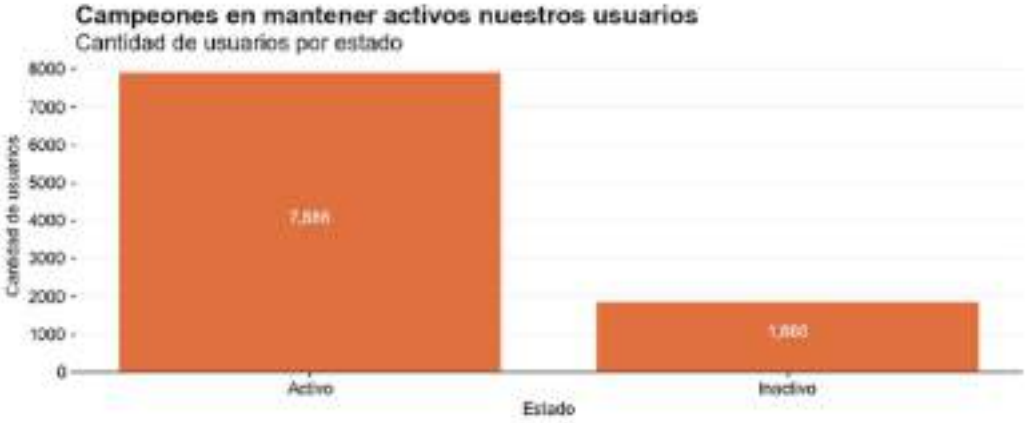
Códigos de marcas

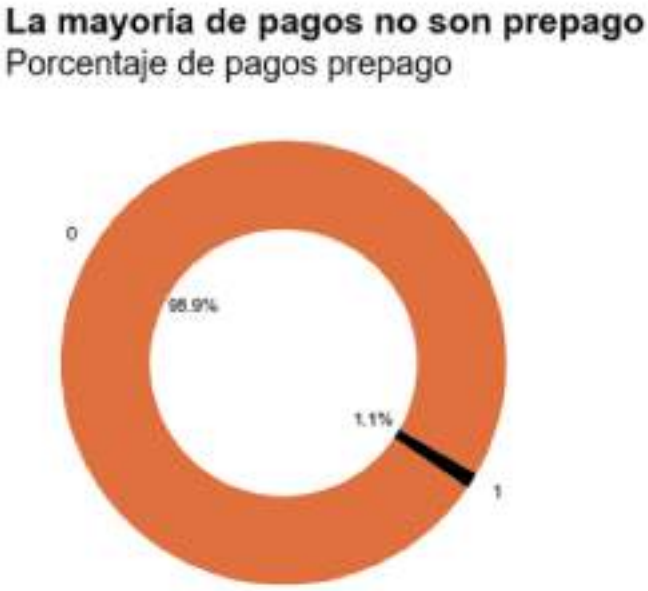
Códigos de compañías

Códigos de organizaciones



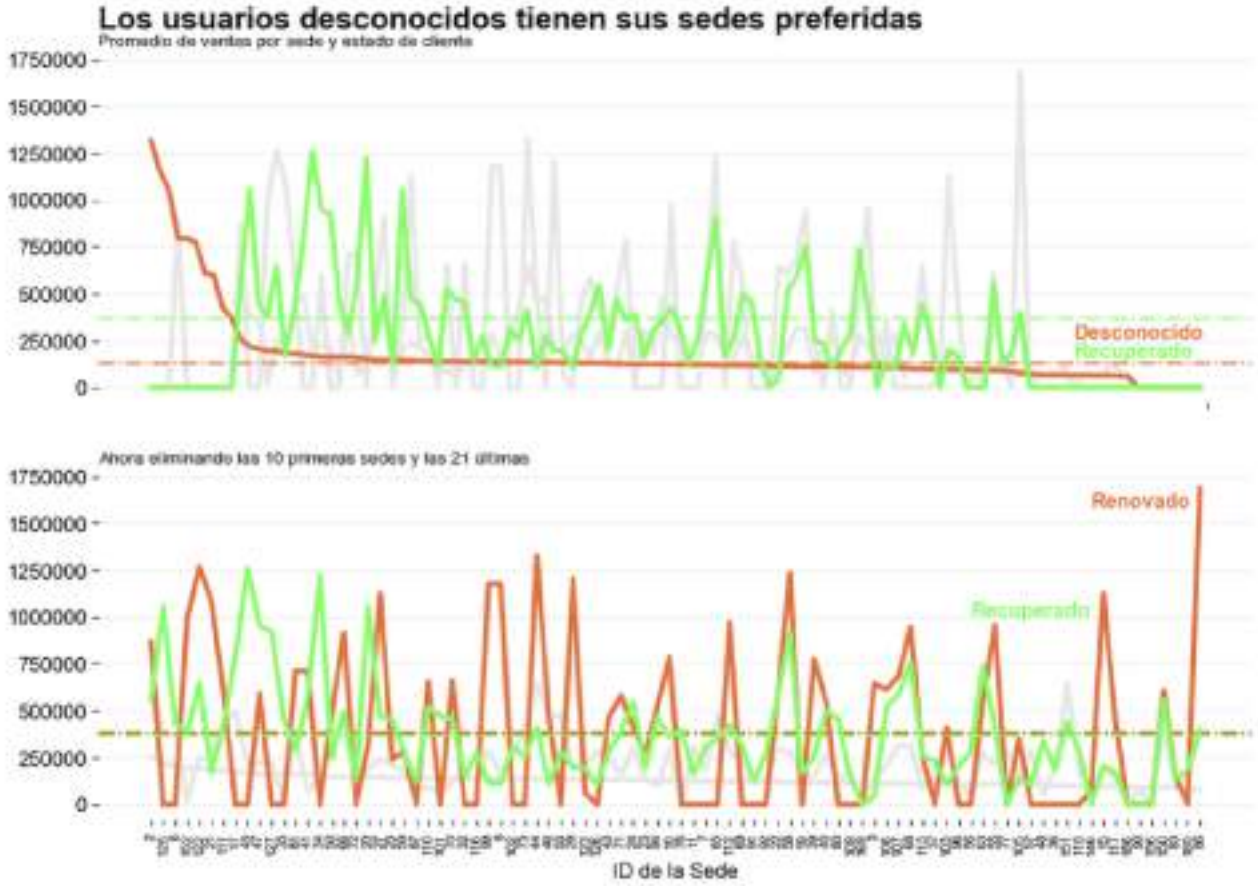
Sería importante conocer más a detalle el venue 156 porque es muy sospechoso la cantidad de ventas que tiene, parece un error en el sistema. Es necesario mencionar que las fechas tienen comportamientos irregulares, hay una venta del 24 de junio, hay 3 después el 8 de julio, el 9 de julio comienzan los registros a las 8 am, pero el 10 se acaba a las 12 del medio día. Revisando los registros de la madrugada, se encuentra muchos débitos, por lo que allí no encontramos mayor inconveniente.





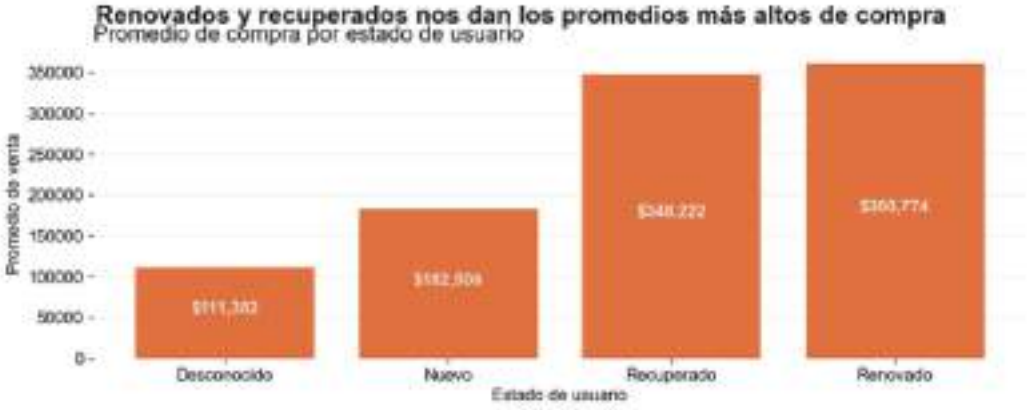
La mayoría de nuestros usuarios que hacen pagos, 7886, están activos, no entendemos cómo hay tantos inactivos a los que les sigue haciendo el débito, no sabemos qué convierte entonces a un usuario en inactivo. Y finalmente, el 98.9% de los pagos no son prepago.

ANÁLISIS MULTIVARIABLE



Es muy interesante ver que cuando quitamos esas sedes en las que no hay muchas ventas de miembros de otros estado, donde los `Desconocidos` tienen un promedio de compra muy elevado, los valores se normalizan un poco de acuerdo a los resultados esperados, donde los usuarios que renuevan tienen un promedio de compra superior a los usuarios nuevos.

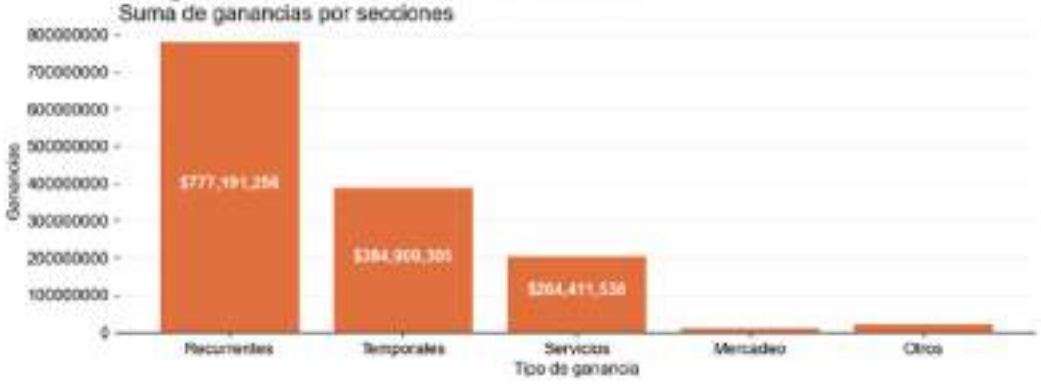
Los ID de las Sedes que se eliminaron son: 86, 131, 133, 134, 135, 136, 137, 138, 139, 140, 141, 144, 145, 147, 148, 149, 153, 155, 157, 159, 160, 161, 162, 163, 164, 165, 168, 171, 183, 184 y 190.



De esta manera podemos concluir que a pesar de tener muchos más miembros desconocidos, los renovados y recuperados son los que más ingreso nos dejan con un promedio de compra de $348.222 COP y $360.774 COP respectivamente.

Las ganancias recurrentes son las más altas

Suma de ganancias por secciones



Los servicios son pocos pero por grandes montos

Histogramas de pagos por tipo de venta

SISTEMA EMPRESARIAL CONFUSO

Product_ID	Products_Prices_ID	Label
35.0	505.0	Afiliación mes VIP
35.0	505.0	Serv Med Deportivo VIP

Es importante trabajar en el formato de los nombres, pues no resulta ser fácil para el análisis, se duplican registros por tildes, incluso hay algunos espacios al final que duplican los registros como en el caso de 'Promo Oct Mes Vip' y 'Promo Oct Mes Vip '. También se hace necesario que `Products_Prices_ID` tenga alguna relación con algún sistema de la empresa, por ejemplo que los valores 3XX representen los servicios como fisioterapia, nutricionista o servicios médicos, mientras los 4XX representan los de mercadeo como activaciones, canjes o cortesías, porque actualmente, por ejemplo, en el código 505 encontramos tanto la `Afiliación Mes Vip` como la `Serv Med Deportivo Vip` que son a $75.000 COP y $49.000 COP respectivamente. El mismo problema ocurre en la columna `Product_ID` como lo evidenciamos anteriormente.

Dado que tenemos tipos de compras diferentes en nuestro dataframe, las filtraremos para hacer más preciso el análisis y lo haremos grupo por grupo para evitar duplicaciones y que todos los labels estén contenidos. Hicimos el filtro manual debido a los problemas que encontramos en las otras columnas, que mencionamos anteriormente.

	Dataframe Name	Shape	Columns	Null_Columns	Final_Columns	Null_Values	Duplicated_Values
0	random_sample_10000-LEADS-tb_crm_leads	(10000, 49)	49	4	45	73444	9911
1	random_sample_10000_Negocios-DEAL-tb_crm_deals	(10000, 82)	82	12	70	95658	7554
2	random_sample_10000_Ventas-Facturas-tb_invoice	(10000, 64)	64	15	49	96194	9850
3	random_sample_10000_Ventas-Facturas-tb_invoice...	(10000, 46)	46	9	37	131849	0

Eliminamos las columnas compuestas únicamente por valores nulos e identificamos que tenemos muchos valores problemáticos tanto nulos como duplicados que debemos revisar. De nuestro primer dataframe se eliminaron 4 columnas, del segundo 12, del tercero 15 y del cuarto 9.

DATAFRAME DE LEADS

Creemos que las columnas actuales son:

- **ID:** Identificador único del registro.
- **UUID:** Identificador único universal (UUID) para el usuario. Lo eliminaremos porque creemos que el ID de Bodytech de la anterior columns es suficiente.
- **Ref Int:** Referencia, aunque actualmente nos encontramos sólo el registro invitado o nulo. Se eliminará porque la columna de CRM recoje esta información.
- **Last Name:** Apellido del usuario. La eliminaremos pues sólo mantendremos el nombre completo y no ser redundantes.
- **First Name:** Primer nombre del usuario. La eliminaremos por la misma razón anterior.
- **Full Name:** Nombre completo del usuario. Vamos a cambiar el formato de los strings para facilitar su lectura.
- **Status:** Estado del usuario, (posiblemente 1 activo / 0 inactivo).
- **Company ID:** Identificador de la empresa, (1 / 3).
- **Organization ID:** Identificador de la empresa (1).
- **Brand ID:** Identificador de la marca, (1 / 2 / 3).
- **Did:** Documento de identidad del usuario. Se eliminará porque aparece de nuevo más adelante.
- **Document Type ID:** Identificador del tipo de documento.
- **Document Number:** Número del documento de identidad.

- **Birth Date:** Fecha de nacimiento del usuario.
- **Genre:** Género del usuario (Masculino/Femenino).
- **Training Goal:** Objetivo de entrenamiento del usuario.
- **Training Level:** Nivel de entrenamiento del usuario.
- **City ID:** Identificador de la ciudad.
- **Venue ID:** Identificador de la sede.
- **Country ID:** Identificador del país.
- **Address:** Dirección del usuario.
- **Latitude:** Latitud de la dirección del usuario. Se eliminará porque sólo hay 0 y valores nulos.
- **Longitude:** Longitud de la dirección del usuario. Se eliminará porque sólo hay 0 y valores nulos.
- **Phone:** Número de teléfono principal del usuario.
- **Phone 2:** Segundo número de teléfono del usuario. Se eliminará porque en la mayoría de los casos es el mismo número de Phone.
- **Email:** Dirección de correo electrónico del usuario.
- **User Creator:** Identificador del usuario que creó el registro.
- **User Last Update:** Identificador del usuario que realizó la última actualización del registro. Se eliminará debido a que es el mismo valor de la columna anterior o un valor nulo, así que no hay información relevante allí.
- **Create At:** Fecha y hora de creación del registro.
- **Create At Db:** Fecha y hora de creación del registro en la base de datos.
- **Update At:** Fecha y hora de la última actualización del registro. Lo eliminaremos porque es el mismo valor que la columna anterior o valores nulos.
- **Contact:** Indica si el usuario acepta ser contactado, pero todos son valores 1, entonces no es relevante, se eliminará.
- **Contact Email:** Indica si el usuario ha optado por recibir correos electrónicos (un valor booleano).

- **Crm Leads Type:** Tipo de lead en el CRM (spontaneous / courtesy / guest / referred)
- **Contact Sms:** Indica si el usuario ha optado por recibir SMS (un valor booleano).
- **Contact App:** Indica si el usuario ha optado por recibir notificaciones a través de la aplicación (un valor booleano).
- **Prospect ID:** Identificador del prospecto.
- **Contact Whatsapp:** Indica si el usuario ha optado por recibir mensajes a través de WhatsApp (un valor booleano).
- **Sales Channel ID:** Identificador del canal de ventas (1 / 3 / 8).
- **Contact Phone:** Indica si el usuario ha optado por recibir llamadas telefónicas (probablemente un valor booleano).
- **Sales Source ID:** Identificador de la fuente de ventas.
- **Members ID:** Identificador de miembros.
- **Photo:** Indica si hay una foto del usuario disponible. Sólo hay la foto de un usuario, se eliminará la columna.
- **Origin Lead:** Origen del lead.
- **Migration:** Indica si el usuario fue migrado de otro sistema. Todos los valores son cero, por tanto se eliminarán.

Únicamente dejamos valores nulos en la columna de `Birth Date` pues no es tan relevante. Donde encuentramos el problema más relevante de todos fue en los 9.911 registros duplicados que no tiene sentido conservar pues es la información del mismo lead, en algunos casos, repetido hasta 110 veces. En realidad, los registros relevantes para nuestro análisis son el 0.89% del dataframe original.

Así que finalizamos nuestra limpieza con un dataframe de 89 registros de leads y 32 columnas que llamaremos `leads.csv`.

DATAFRAME DE DEALS

Además de las columnas del dataframe anterior y que ya eliminamos tenemos las siguientes:

- **Total:** No se entiende el contexto y todos los registros tienen el valor 0.
- **Position:** La posición del embudo en el que se encuentra el usuario, pero todos los registros son 0.
- **Status:** El estado del usuario, pero es *True* en todos los registros.
- **Sell Assigned Status:** Es 1 para todos los registros.
- **Sales Segment ID:** Es 4 para todos los registros.

Volvimos a dejar valor nulos únicamente en la fecha de nacimiento, pues no parece ser relevante ahora. De nuevo el problema más importante estuvo en la cantidad de duplicados que tiene nuestro dataset, en este caso, tenemos hasta 341 registros idénticos que no tiene ningún sentido conservar, así que los eliminamos.

Finalmente, nuestro dataframe de 2446 registros únicos lo guardamos en un nuevo archivo csv llamado `deals.csv`.

DATAFRAME DE SELLS

En este dataframe observamos las siguientes columnas:

- **ID:** Identificador único de la factura.
- **Ref:** Referencia interna de la factura.
- **Ref Ext:** Referencia externa de la factura. Resultó ser la misma interna, por tanto se elimina.
- **Consecutive:** Número consecutivo de la factura.
- **Customer ID:** Identificador único del cliente.
- **Brand ID:** Identificador de la marca.
- **Company ID:** Identificador de la compañia.
- **Organization ID:** Identificador de la organización. Es 1 en todos los registros, por tanto se elimina.
- **Venue ID:** Sede donde se realizó la venta.
- **Users Creator:** Identificador del usuario que creó la factura.
- **Users Close:** Identificador del usuario que cerró la factura. Usualmente el mismo que creó la factura o nulo, por tanto se elimina.
- **Status:** Estado de la factura. Todos los valores son True, así que se elimina.
- **Status ID:** Identificador del estado de la factura. Todo los valores son validated, por tanto se elimina.
- **Total Sub:** Total sin impuestos (sub-total) de la factura.
- **Discount Percent (COP):** Porcentaje de descuento aplicado en COP (Peso Colombiano). Sólo hay 0, 100 o nulos, es necesario formatearla.
- **Invoice Status 2:** Estado adicional de la factura (full_payment / validated).
- **Discount Absolute (COP):** Descuento absoluto en COP.

- **Discount Absolute (COP):** Descuento absoluto en COP.
- **Discount (COP):** Descuento total en COP.
- **Total Ht:** Total antes de impuestos.
- **Tva:** Impuesto sobre el valor agregado (IVA).
- **Localtax1:** Impuesto local 1. Todos los valores son 0 o nulo, por tanto se elimina.
- **Localtax2:** Impuesto local 2. Todos los valores son 0 o nulo, por tanto se elimina.
- **Total Ttc:** Total incluyendo todos los impuestos.
- **Manual Discount:** Descuento aplicado manualmente. Todos los valores son False, así que se elimina.
- **Created At:** Fecha y hora de creación de la factura.
- **Create At Db:** Fecha y hora de creación en la base de datos. Usualmente el mismo registro de la columna anterior, por tanto se elimina.
- **Update At:** Fecha y hora de la última actualización. Usualmente con diferencias de segundos con respecto a la columna anterior que podrían ignorarse, entonces se elimina.
- **Sales Channel ID:** Identificador del canal de ventas.
- **Sales Segment ID:** Identificador del segmento de ventas.
- **Note Private:** Nota privada asociada a la factura.
- **Note Public:** Nota pública asociada a la factura.
- **Due Date:** Fecha de vencimiento de la factura.
- **Agreement ID:** Identificador del acuerdo asociado con la factura.
- **Proposals ID:** Identificador de la propuesta asociada con la factura.
- **Invoice Type:** Tipo de factura ('FAC' para factura / BLT001 / NOTA_CREDITO).
- **Invoice Billing Ruling ID:** Identificador de la normativa de facturación.
- **Assigned Executive:** Ejecutivo asignado a la factura.
- **Assigned Executive Secundary:** Ejecutivo secundario asignado a la factura. Usualmente nulo, por tanto se elimina.

- **Ref Consecutive:** Referencia consecutiva adicional. Igual a las referencias anteriores, entonces se elimina.
- **Invoice ID:** Identificador de la factura.
- **Pdf:** URL del archivo PDF de la factura.
- **Status Send Gp:** Estado del envío a GP (posiblemente sistema de gestión) (pending / success).
- **Status Send Electronic:** Estado del envío electrónico. Todos los valores son success, por tanto lo eliminamos.
- **Credit Note Type:** Tipo de nota de crédito.
- **Migration:** Información relacionada con la migración de datos. Todos los valores son 0, por tanto se elimina.
- **Corporate Invoice Order ID:** Identificador del pedido de factura corporativa.
- **Crm Is New:** Indica si el CRM es nuevo.
- **Event:** Evento asociado con la factura.
- **Cufe:** Código Único de Factura Electrónica.

En el caso de este tercer dataframe si nos aseguramos que no quedaran valores nulos, todos fueron completados. Pero, lastimosamente, vuelve a ser alarmante la exagerada cantidad de registros duplicados, 9850 en total que fueron eliminados. Así que guardamos nuestro dataframe final de 150 filas por 36 columnas con el nombre `sells.csv`.

DATAFRAME DE INVOICES

Suponemos que estas son nuestras columnas:

- **ID:** Identificador único del registro del producto en la factura.
- **Product ID:** Identificador único del producto.
- **Products Prices ID:** Identificador del precio del producto.
- **Label:** Etiqueta del producto, nombre.
- **Description:** Descripción del producto.
- **Total Sub:** Total sin impuestos (sub-total) del producto.
- **Discount Percent (COP):** Porcentaje de descuento aplicado en COP (Peso Colombiano).
- **Discount Absolute (COP):** Descuento absoluto en COP.
- **Discount (COP):** Descuento total en COP.
- **Total Ht:** Total antes de impuestos del producto.
- **Tva:** Impuesto sobre el valor agregado (IVA) aplicado al producto.
- **Totaltax1:** Primer impuesto local aplicado al producto. Todos los valores son 0 o nulos, por tanto se elimina.
- **Totaltax2:** Segundo impuesto local aplicado al producto. También se elimina porque todos los valores son 0 o nulos.
- **Total Ttc:** Total incluyendo todos los impuestos para el producto.
- **Venue ID:** Identificador del lugar donde se vendió el producto.
- **Create At:** Fecha y hora de creación del registro del producto. Contiene muchos valores nulos y los no nulos son iguales a los de la siguiente columna, por tanto se elimina.

- **Update At Db:** Fecha y hora de la última actualización en la base de datos. Se eliminará porque parece duplicada de la columna anterior.
- **Users ID:** Identificador del usuario asociado al registro del producto.
- **Assigned Executive Secundary:** Ejecutivo secundario asignado al registro del producto. Deberíamos tener un principal, pero no lo tenemos.
- **Invoice ID:** Identificador de la factura a la que pertenece el producto.
- **Promo ID:** Identificador de la promoción aplicada al producto.
- **Agreement Line Deferred Payment ID:** Identificador de la línea de acuerdo para el pago diferido del producto.
- **Promotion Price ID:** Identificador del precio promocional aplicado al producto.
- **Ruler Member Status ID:** Identificador del estado del miembro en el sistema de reglas (nuevo / recuperado / renovado).
- **Migration:** Información relacionada con la migración de datos del producto. Todos los valores son 0 y uno nulo, se elimina la columna.
- **Promotion Price Mes ID:** Identificador del precio promocional mensual del producto.
- **Ruler Product Status ID:** Identificador del estado del producto en el sistema de reglas.
- **Is Commissionable:** Indica si el producto es comisionable. Todos los valores son 0 y uno nulo, se elimina la columna.
- **Brand ID:** Identificador de la marca del producto.
- **Company ID:** Identificador de la compañía.
- **Organization ID:** Identificador de la organización.
- **Nc Invoice ID:** Identificador de la nota de crédito asociada con el producto. Sólo hay un valor no nulo, se elimina la columna.
- **Status:** Estado del registro del producto.
- **Venue Use:** Uso del lugar asociado con el producto.
- **Is Prepaid:** Indica si el producto ha sido prepagado.

Este resultó ser el dataframe más completo de todos, pues realmente terminamos con 9.999 registros para trabajar correctamente y 29 columnas realmente significativas. Únicamente hubo un registro que tenía todos los valores nulos y sólo un número de factura que eliminamos por los problemas que nos podría representar.

Finalmente guardamos el dataframe final como un archivo csv con el nombre `invoices.csv`.

CONCLUSIONES

Luego de analizar el dataframe de las facturas podemos decir respecto a las ganancias que:

- Encontramos muchos valores atípicos en la columna de Total_Sub con los subtotales a pagar.
- Por lo general, Bodytech no aplica descuentos.
- La sede con el ID 156 tiene 1300% la cantidad de ventas de la sede 36 que es la segunda en cantidad de ventas.
- La mayoría de las ventas corresponden a los días 9 y 10 de julio de 2024. Sin embargo, los días en que mayor promedio de ventas hubo fueron el 24 de junio y el 8 de julio de 2024.
- El 77% de los pagos provienen de usuarios desconocidos, el 13.3% de usuarios nuevos, el 7.2% de usuarios recuperados y el 2.5% de usuarios renovados.
- El dataframe no es claro al categorizar las marcas, compañías y organizaciones.
- La mayoría, 7.886 son usuarios activos, sin embargo 1.860 usuarios inactivos a los que se les sigue aplicando el débito, entonces ¿qué significa inactivo para la empresa?
- El 98.9% de los pagos son prepago.
- Existen 31 sedes en las que no hay muchos pagos, dentro de estas, hay 10 que son las preferidas para los usuarios Desconocidos.

- El promedio de compra más alto, de $360.774 COP nos lo dan los usuarios `Renovados`, luego, con un promedio de $348.222 COP están los usuarios `Recuperados`.
- Es importante revisar la forma en la que el sistema está guardando la información de las columnas `Product_ID`, `Product_Prices_ID` y `Label` debido a que hay muchos problemas de formato y de incoherencia, porque hay varios registros que comparten códigos pero no tienen ninguna relación en común. Esa es nuestra propuesta, que de acuerdo al tipo de servicio tenga un dígito inicial.
- Las que denominamos ganancias recurrentes son los pagos donde más dinero nos ingresa con $777'191.256 COP, seguido de los pagos temporales con $384'900.305 COP.
- Las que denominamos ganancias por servicios no genera muchos pagos, pero son los de más alto valor.

BODY TECH®

GRACIAS

ELABORADO POR:
JUAN PABLO RAMOS BEDOYA