Team Origin
Atmira Pharma Visualization

Reto Atida Mifarma

Para realizar este proyecto vamos a seguir las siguientes fases de un proceso de analítica de datos: Entender el reto, Adquisición de datos, Limpieza de datos, Análisis, Visualización y Acción.

Las herramientas que hemos utilizado en el proyecto son:

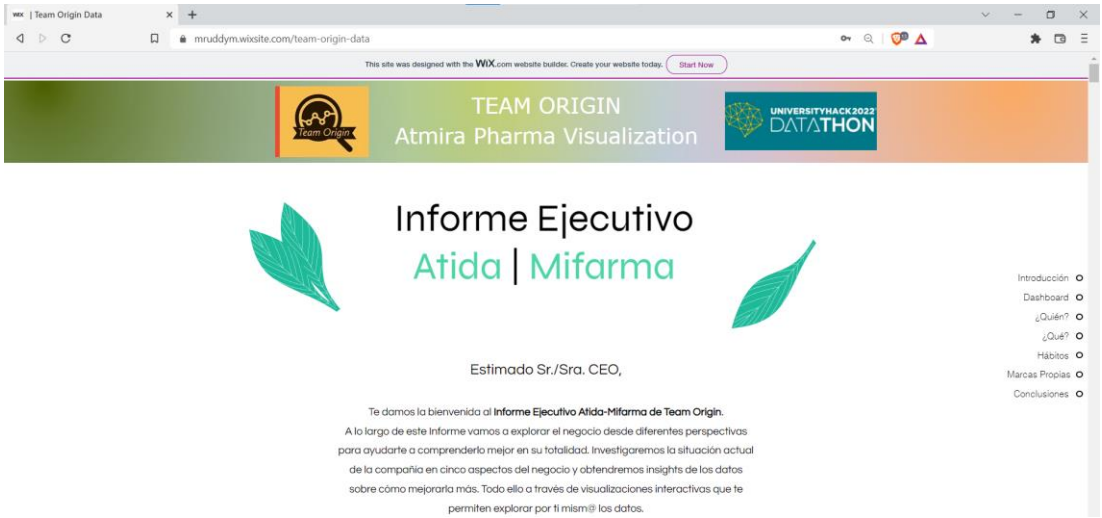
- **Jupyter Notebook** (Python): para estructurar y ordenar todo el código
- **Pandas** (Python): limpieza y análisis de datos
- **SQLite3** (Python): análisis de datos en lenguaje SQL
- **Google Sheets:** para cálculos y para guardar los datos online de forma privada
- **Google Data Studio:** para visualizar los datos guardados en Sheets
- **Wix.com:** para construir y hostear la página web con las visualizaciones insertadas

1. Entendiendo el reto

Vamos a realizar un informe ejecutivo del que ayude a los directivos de Atida a comprender mejor su negocio y a tomar mejores decisiones basadas en los patrones y oportunidades identificadas. Para ello vamos a comparar el año 2018 con el 2017 a nivel general y planteando una serie de preguntas dirigidas a entender mejor algunos aspectos concretos del negocio.

El informe explica mediante gráficos interactivos cómo es el negocio. Muchos aspectos de los gráficos son ajustables para permitir una experiencia interactiva al usuario. En base a los resultados, se extraen conclusiones y se propone sugerencias de Toma de Decisiones que ayuden a mejorar el negocio.

El formato es de una web funcional, hosteada con privacidad en wixsites.com. **Se requiere una contraseña para acceder**, que es datathon22



Muestra de la página web. URL: https://mruddym.wixsite.com/team-origin-data

2. Adquisición de datos

El informe utiliza como fuente principal los datos facilitados por Atida, debidamente limpiados. Estos incluyen los datos de ventas de 2017 y 2018 y una clasificación de los productos vendidos. Los hemos enriquecido con datos de otras fuentes:

- **INE:** datos de demografía de España. Datos de códigos de Comunidades y Provincias. Para estudiar la distribución geográfica de las ventas y la demografía de los clientes.
- **Google APIs:** datos de localización geográfica de municipios, códigos postales y regiones. Para crear mapas por provincia, comunidad y municipio.
- **Agencia Tributaria:** datos de Renta en ciudades por código postal. Para aproximar el poder adquisitivo de los clientes de las ciudades.
- **Web Atida:** lista de marcas propias. Para hacer un estudio particular de las marcas propias.

3. Limpieza de datos

1. Limpieza de datos facilitados por Atida

En los ficheros *products.csv* y *products_categories.csv* hemos cambiado algunos caracteres que daban problemas al descodificarlos, como ñ y los acentos. En *products.csv* hemos visto que algunos "product_id" estaban repetidos. Tras estudiarlo, se trata de pequeñas variaciones en la descripción del producto, tales como cambiar el orden de las palabras, pero manteniendo otras variables como "sku" idénticas. Para evitar duplicados en fases posteriores del análisis, hemos decidido eliminar las entradas repetidas, con el criterio de mantener la primera que apareciera.

En el fichero *items_ordered_2years.txt* hemos realizado varias fases de limpieza, sobre todo en función de los parámetros zipcode y city, los cuales presentan múltiples tipos de errores. Entre ellos, se encuentran: incluir códigos postales de otros países como Portugal, China e Inglaterra; tener invertidos los campos zipcode y city; incluir códigos postales con "o" u "O" en vez de 0; incluir un espacio al principio o final del campo zipcode; entre muchos otros. Los más comunes han sido detectados y corregidos en caso de ser códigos españoles, mediante funciones y algunos a mano. En el caso de los códigos de otros países, o de aquellos ausentes, hemos decidido eliminarlos del dataset para limitar el alcance del análisis a productos vendidos en España de los que conocemos el código postal.

A continuación, se indican las cifras del proceso de limpieza de *items_ordered_2years.txt*:

Entradas iniciales	**930914**
Entradas correctas	+ 898897
Códigos otros países	- 24227
Otras entradas eliminadas	- 6679
Entradas subsanadas	+ 3537
Entradas finales utilizadas	**= 900008 (97%)**

Hemos realizado esta limpieza utilizando la librería Pandas de Python. A los datos iniciales hemos añadido dos columnas "weekday" y "provincia", con el nombre del día de la semana de la compra (lunes, martes…) y el nombre de la provincia de la compra (Álava, Albacete…), para facilitar el análisis posterior.

También hemos solucionado algunos errores de la columna "base_cost" en los que este campo era excesivamente grande en comparación a "price" (por ejemplo, "base_cost" = 54000,00€ y "price" = 63,00€) lo cual generaba errores al calcular el beneficio.

Los detalles de cómo se han limpiado todos los datos se pueden ver en el fichero de Jupyter Notebook *limpieza.ipynb*.

Hemos preparado los datos de las otras fuentes con Google Sheets y un editor de texto. Su proceso ha sido sencillo y poco relevante.

2. Preparar base de datos

Para poder analizar todos los datos, hemos creado una base de datos *atida.db* que operar con SQLite desde Python (librería SQLite3). SQLite es una versión gratuita de SQL que no requiere de servidor y permite hacer consultas ("queries") muy flexibles para hacer análisis de los datos. Cada uno de los ficheros facilitados por Atida, una vez limpios, los hemos introducido como una tabla. Este proceso está implementado en *crear_database.ipynb*. Posteriormente hemos añadido, para otras partes del análisis, una tabla con todos los clientes y otra con todos los productos de Marca Propia de Atida-Mifarma.

4. Análisis

Mediante (muchas) consultas de SQLite que se pueden ver en *analisis.ipynb*, hemos ido obteniendo nuevas tablas con los resultados relevantes para las visualizaciones. Estas son las tablas que posteriormente hemos subido de forma privada a Google Sheets. Desde allí son accedidas por Google Data Studio para crear las visualizaciones.

Para clasificar los productos por categorías, hemos utilizado principalmente la variable *analytic_category* del fichero *products.csv* que divide los items en:

Cosmética y Belleza	308218	Otros	89297	Ortopedia	8546
Higiene	213649	Nutrición	38934	Perfumería	1385
Infantil	119593	Vida Íntima	16926	Veterinaria	254
Herbolario	90512	Óptica	12694	TOTAL	900008

Para algunas visualizaciones hemos utilizado la clasificación en *cat1*, *cat2* y *cat3* del fichero *products_categories.csv* porque permite una mayor precisión al tener más sub-categorías. El lado negativo de esta clasificación es que incluye menos productos, y por tanto hay más que aparecen con categoría nula. Por eso hemos decidido utilizar, en lo posible, la primera clasificación.

Para el análisis, hemos incluido los siguientes apartados:

1. Dashboard General
    - KPIs de Ventas, evolución de ingresos y ventas por categoría
2. ¿Quiénes son nuestros clientes?
    - Distribución geográfica
    - Penetración de mercado. Demografía
3. ¿Qué compran nuestros clientes?
    - Marcas y productos más populares
    - Paquetización de productos y categorías
4. Entendiendo los hábitos de compra
    - Distribución temporal de ventas
    - Generación de ventas recurrentes
    - Captación de clientes nuevos

Vista del Dashboard General

5. Marcas Propias
    - Análisis de marcas propias
6. Recomendaciones finales

Para cada uno de ellos, hemos hecho un análisis a partir de los datos para ayudar a responder a las preguntas, buscando una mejor comprensión del negocio e identificando oportunidades de crecimiento. Al final de cada sección, ofrecemos un resumen con las conclusiones del análisis, indicando los hallazgos y sugiriendo cómo estos se pueden usar para el beneficio de la empresa.

5. Visualización

El informe está hecho de tal forma que el usuario puede explorar las visualizaciones de forma interactiva y llegar por sí mismo a las mismas observaciones que están expuestas. Las visualizaciones, por tanto, están pensadas como herramienta para demostrar las conclusiones y a la vez permitir una exploración de los datos. Cada parte del análisis tiene sus correspondientes visualizaciones.

Se incluyen visualizaciones de varios tipos. En primer lugar, están los Dashboards, que agrupan varios gráficos sobre un mismo tema, junto con KPIs (Key Point Indicators). También incluye mapas, tablas y diagramas de treemap, todos interactivos. La interactividad se basa en poder elegir uno o varios parámetros del gráfico, como la fecha o la categoría del producto, y que el gráfico se recalcule para la selección hecha. También permite poner el cursor sobre los elementos del gráfico para ver su nombre y valor.

Para visualizar los datos, hemos creado visualizaciones de Google Data Studio, una herramienta gratuita que permite adjuntarlos a páginas web. Data Studio obtiene los datos en directo de Google Sheets, donde los hemos subido.

6. Acción

El análisis y las visualizaciones no aportan nada si no se toman decisiones en base a ellos. Para motivar la acción, al final de cada apartado hemos incluido sugerencias de toma de decisiones, en base a las observaciones del análisis que hemos realizado. Están dirigidas a los directivos de Atida y comunican, de forma breve, las pautas clave de cada sección. Además, al final del Informe se incluye un apartado de conclusiones, donde se indican las zonas de expansión identificadas, se sugieren rutas a seguir y se indican algunas amenazas detectadas.

¿Cómo vamos a mejorar el proyecto en el futuro?

Vamos a profundizar en el análisis de paquetización de productos y de captación de clientes nuevos, para ofrecer conclusiones y recomendaciones más concretas que sean inmediatamente aplicables. Además, vamos a incluir al principio del informe un resumen con los hallazgos más importantes, de forma resumida, para crear impacto en el lector y captar su atención. También vamos a añadir otras herramientas de visualización, como Datawrapper, que incluyen gráficos con un mejor diseño visual que Data Studio, para que el informe sea más impactante visualmente. Es posible que añadamos alguna sección más al informe para complementar el análisis actual.