**ANÁLISIS INGRESOS ENTRE TIENDAS DE CAFÉ:**

**Paso a paso como he hecho todo y como he razonado todo**

1. Contexto y objetivo del análisis:

* Se trata de 3 tiendas de café ubicadas en distintas localizaciones (Lower Manhattan, Hell’s Kitchen y Astoria) abiertas recientemente (hace 6 meses) y este análisis se hace con el propósito de poder observar el rendimiento de las 3 tiendas en estos 6 primeros meses de apertura, tratando de buscar si ha habido diferencia significativa en los ingresos recaudados por cada una de ellas, y si es así, analizar el por qué en esas diferencias en relación con cómo operan cada una de ellas. Este primer análisis es importante ya que la identificación de diferencias entre ellas puede afectar a que cuando lleven más tiempo en el mercado, esas diferencias se hagan aún más significativas y desde un primer momento está bien identificarlas para trabajar sobre ellas y poner hincapié en aquellos aspectos que sean necesarios de forma que ninguna de las 3 se vea con problemas de ingresos ni ahora ni más adelante por conductas que se pueden corregirse.

2. Analizar las columnas de las que dispongo:

* Transaction\_id: esta columna es el identificador único de la tabla. Cada fila representa una transacción y cada transacción recibe un valor único en esta columna. Es una columna que lógicamente tiene que permanecer en el análisis y será útil para hacer recuentos de transacciones. De primeras, podríamos decir que contamos con 149.119 transacciones, pues es el número de filas que contiene la tabla. Pero de un simple vistazo, se observan dos cosas: la primera, que al final de la tabla parece que hay dos duplicados que más adelante habrá que tratar. La segunda, obviando esos dos duplicados, el número de fila del Excel no coincide con el número de identificador único al final de la tabla, algo que habrá que observar por qué. Se cambia el formato a número.
* Transaction\_date: lo primero que hago es cambiar el formato a fecha. Observo que hay ahí unos valores que no se han cambiado correctamente al formato fecha y ocurre porque están almacenados como texto y no como número. Los convierto primero a número y después a fecha. Esta columna indica el momento (día/ mes/ hora) en el que ocurre la transacción y me parece una columna relevante en el análisis, pues puede ser muy interesante analizar la actividad de las 3 tiendas a lo largo de los 6 meses o a lo largo de los días de cada mes, y ver si es constante, si se da un pico de ventas en unos meses o unos días determinados del mes etc. El año es más irrelevante por ahora, pues todo transcurre en el 2023.
* Transaction\_time: lo primero que hago es cambiar el formato a hora. Esta columna indica la hora exacta en la que ocurre la transacción. También considero que es una columna relevante para el análisis ya que observar para cada tienda, en qué momentos del día vende más o menos puede ser relevante a la hora de también luego llevar a cabo estrategias concretas etc.
* Transaction\_qty: la cambio a formato número. Nos informa sobre la cantidad del producto x que se ha pedido en cada transacción. También es una columna relevante en este análisis. Que una tienda venda mayor o menor cantidad de productos en cada transacción puede tener mucho que ver con los ingresos que genera. Y de nuevo, como todas las columnas prácticamente, puede ser también muy interesante para a posteriori elaborar estrategias concretas para cada tienda.
* Store\_id: por lo que observo esta columna es el identificador único de las tiendas (lo observo en la siguiente columna de store\_location). Filtrando, el 3 corresponde a Astoria, el 5 a Lower Manhattan y el 8 Hell’s Kitchen. Por comprobar que esto es así en todas las filas y no hay ningún error, me he generado una tabla dinámica en la hoja “Comprobaciones” que me lo confirma. Ya en esta tabla dinámica puedo observar también que hay valores nulos en la columna de store\_location (aunque no entiendo por qué no me los cuenta), y que tienen asociado un store\_id de 5 por lo que son nulos fáciles de tratar ya que se pueden rellenar, en este caso hay que rellenarlos con “Lower Manhattan”. He decidido que esta columna de store\_id la voy a ocultar, pues teniendo ya la de store\_location me parece redundante tenerla. En caso de estar haciendo un análisis con SQL podría ser interesante mantenerla, pero al ser con Excel no.
* Store\_location: lo primero que hago es ponerla en formato de texto. Esta columna es clave en el análisis, pues con ella crearé todas las tablas dinámicas etc. para poder estudiar por tiendas el resto de las variables. Como ya previamente he identificado que hay nulos y son sencillos de tratar, voy a tratarlos en este mismo momento. Filtro quedándome con las celdas vacías, observo que solo hay 3 nulos que son los que se observan de primeras en la tabla, pues están localizados en las primeras filas. Aunque son muy pocos registros, en la hoja “Comprobaciones” queda demostrado que estos nulos suponen un 0,002%, por lo que nada relevante. Simplemente relleno las 3 celdas vacías con “Lower Manhattan”.
* Product\_id: con esta columna ocurre parecido a con la de store\_id. Filtrando, puedo observar que esta columna es el identificador único de cada producto completo, cada producto específico. Es decir, cada combinación de product\_category, product\_type y product\_detail tiene asociada un valor único en product\_id. De nuevo, para comprobar que esto es así en todos los registros y que está todo correcto, genero una tabla dinámica en la hoja “Comprobaciones” que me lo afirma. En esta tabla observo que en product\_category hay algunas celdas vacías. Hay una que corresponde al product\_id 22, otra al 28 y otra al 77. De la misma forma que con store\_id he decidido ocultar esta columna.
* Unit\_price: Esta columna informa sobre el precio unitario del producto en cada transacción. Sé que es el precio unitario de cada producto porque filtrando por cualquiera de ellos, se observa que corresponde a un mismo producto y que además es el mismo para todas las facturas (tengan más o menos cantidad del producto). Es imprescindible para obtener los ingresos de cada tienda ya que esta columna se va a utilizar para calcular el total\_price que es el valor que realmente nos interesa (generando una nueva columna) y que se obtendrá como el producto entre unit\_price y transaction\_qty. Esta columna de total\_price la convierto a formato moneda y la columna de unit\_price la voy a mantener también porque me interesa poder analizar en qué cantidad las tiendas venden productos de precios más bajos, o altos…
* Product\_category: la pongo en formato texto. Esta columna indica a que categoría pertenece el producto ordenado en cada factura. Obviamente esta columna es imprescindible en el análisis también. Dependiendo de la calidad de tus productos, el estado en el que los vendas etc. puede ocurrir que una tienda venda muchos más productos de una categoría concreta a otra, por ejemplo. Interesa estudiarlo también. Con product\_id ya había identificado tres celdas vacías que eran fáciles de rellenar gracias a disponer del product\_id. Lo primero que hago es, filtrando, comprobar que eran efectivamente 3 celdas vacías y a continuación relleno con sus correspondientes categorías. Filtrando con product\_id, la categoría correspondiente al 77 es “Bakery”, al 22 es “COFFEE” y al 28 es “COFFEE” también. Las 3 celdas vacías quedan ya rellenas con sus correspondientes valores. En este caso, el % de nulos es el mismo que en store\_location con lo que es irrelevante. En esta columna, a simple vista también hay algo a corregir y son las mayúsculas y minúsculas. Lo que hago es reemplazar “COFFEE” por “Coffee”, “COFFEE BEANS” por “Coffee Beans” y “tea” por “Tea”, de forma que quede todo igual (la primera letra de cada palabra en mayúsculas y el resto en minúsculas).
* Product\_type: la pongo en formato texto. Esta columna también decido mantenerla, pues en vez de estudiar por product\_category puede interesar ir un poco más allá, profundizando más en los productos y estudiar por product\_type. En cuanto a las mayúsculas y minúsculas en esta columna ocurre un poco lo mismo y también tengo que realizar varios reemplazos (por ejemplo: “Green beans” por “Green Beans”)
* Product\_detail: esta columna la voy a ocultar, pero antes de nada la voy a utilizar para haber podido ocultar la columna de product\_id sin perder la información de esa columna. Es decir, la columna de product\_id lo que hacía era darnos un identificador único sobre los productos específicos. Disponemos de las columnas por separado (product\_category, product\_type, product\_detail) pero no de una que sea como la correspondiente a la de product\_id, y puede interesar tenerla para poder analizar los productos no solo por separado en cuanto a categorías y tipos sino en cuanto al detalle total del producto, al producto específico, disponiendo de los 3 valores a la vez (categoría, producto y detalle). Para ello lo que he hecho es generarme una nueva columna que es product\_specific concatenando las 3 columnas. Disponiendo de esta nueva columna, la de product\_detail ya es redundante tenerla. Antes de generar esta nueva columna corrijo mayúsculas y minúsculas también en product\_detail (por ejemplo: “Carmel syrup” a “Carmel Syrup”). Una vez ahora dispongo de esta columna de product\_specific lo que voy a comprobar es que cada producto específico tenga asociado un único unit\_price para comprobar que no hay errores o ningún valor extraño en la hoja “Comprobaciones”. Para mi sorpresa, me encuentro con que hay varios productos los cuales tienen asociados más de un precio unitario, por lo que es algo a observar el por qué.

En este punto ya tengo dos cosas que investigar: por un lado, los valores de transaction\_id y por otro lado la diferencia en los unit\_price de los productos.

* En cuanto al problema con transaction\_id, lo pude identificar un poco a simple vista. Voy observando por tramos e identifico que el primer salto ocurre cuando el ID pasa de 3251 a 3281. El segundo cuando pasa de 20650 a 20678. En este punto ya me doy cuenta de que está ocurriendo en los días 6 y 7 de cada mes, y, efectivamente, en el salto del día 6 al 7 en los 6 meses hay un salto en el ID de las facturas, de forma que se están perdiendo algunas. Se notifica para que sea revisado y ver que está ocurriendo.
* En cuanto al precio unitario, tras hacer varias comprobaciones e identificar que no tiene que ver con los meses, ni con las horas del día, logro identificar que sí hay un patrón en los días del mes (hoja “Comprobaciones). Esta diferencia en el precio de los productos ocurre los días 9, 30 y 31, especialmente en el 9 (aunque no ocurre en todos los meses), y principalmente se da en la tienda de Lower\_Manhattan, aunque también en las otras. Esto también queda notificado para ser observado el por qué.

3. Comprobar si verdaderamente hay diferencias en los ingresos de las tiendas:

* Esta comprobación la hago en la hoja “Ingresos”. Ocurre que los ingresos de las 3 tiendas a lo largo de estos 6 meses han sido muy similares. Todos rondan los 23.000 euros y tan solo hay una diferencia de 6.500 euros aproximadamente entre la que más ha recaudado (Hell’s Kitchen) y la que menos (Lower Manhattan). Aún así, aunque la diferencia en los ingresos sea pequeña este análisis puede ser igualmente útil de cara a un futuro, pues si se logra identificar alguna diferencia en la conducta de las tiendas que explique esa pequeña diferencia en los ingresos, puede ser trabajado a tiempo para que esta diferencia no se haga cada vez más grande.

4. Creación de columnas nuevas identificadas al hacer el EDA:

Sugerencias: Una columna que identifique si hay varias transacciones que corresponden a la misma factura. Otra columna de customer\_id que identifique que transacciones corresponden a un mismo cliente.

Transaction\_day: al igual que interesa estudiar la actividad de las tiendas por mes, también me gustaría estudiarlo por días del mes, y como con la columna de fecha completa no me deja agrupar por días del mes en la tabla dinámica lo que he hecho es crearme una nueva columna con únicamente los días del mes de forma que pueda realizar esta agrupación correctamente

Para el estudio de las ventas por horas del día me planteé crear una columna nueva para hacerlo por rangos (mañana /tarde/ noche) pero al final decidí que no ya que tampoco había demasiadas horas como para que fuese un análisis complicado de observar y además me parece más interesante conocer la hora exacta a la que una tienda vende más o menos (me parece más útil y de cara a luego poder establecer estrategias etc. también).

5. Tratamiento de nulos

En esta base de datos concreta los nulos han sido tratados previo al EDA ya que en este caso eran nulos fácilmente identificables pues eran pocas celdas vacías en únicamente dos columnas, y, además, muy fáciles de rellenar. Al hacer el EDA no se han identificado más nulos a tratar ni nada, pero por practicar y asegurar he calculado el porcentaje de nulos en cada columna, confirmando que hay 0% en todas (los pocos que había ya se rellenaron y suponían en ambas columnas el 0,002%).

6. Tratamiento de duplicados

Para esta base de datos el tratamiento de duplicados resulta también muy sencillo. Como yo todavía no puedo hacer uso de power query lo que he hecho es contar cuantos registros únicos (tomando como referencia transaction\_id) hay y la diferencia entre el número de registros y los registros únicos son los duplicados. En este caso solo hay 2 y además son dos fácilmente identificables, pues están al final de la tabla y ya en los análisis los había observado. Como solo cuento con estos dos suplicados, sin más complicación, los he pegado en una hoja para poder identificar si todos los campos eran exactamente iguales o no, y efectivamente eran todos iguales, por lo que los he eliminado de la tabla y ya.

7. Resultados para el dashboard:

Que KPIS calculo para ponerlos en el dashboard

En general no se han observado diferencias significativas en la actividad y la forma en la que operan las 3 tiendas, pero hay dos aspectos que cabe destacar pues, aunque la diferencia no sea muy grande, la hay: las horas del día a las que se vende más o menos, la cantidad de productos que vende cada una. Se ha observado que Lower Manhattan es la única que logra vender determinadas categorías de productos en cantidades de 3, de donde puede sacar provecho y tener una ventaja competitiva. Por otro lado, Lower Manhattan es la que

Al hacer el análisis si he observado dos cosas a destacar también a la empresa para tener en cuenta, una más bien positiva y otra más bien negativa:

* Hay productos que no se están vendiendo prácticamente nada
* Las ventas han ido incrementándose con el paso de los meses

Se podría dar alguna sugerencia sobre que columnas como la edad de los clientes que compran pueden ser interesantes a la hora de establecer estrategias, o su género