LABORATORIO EXCEL

1. LIMPIEZA DE DATOS Y TRANSFORMACIÓN DE DATOS

Observando los datos, una de las primeras cosas que he visto es, al final de los datos dos transacciones duplicadas. Al ser el id el mismo en un principio podría pensar que corresponde a la misma transacción pero que es un desglose de dos productos diferentes que se han comprado en esa misma factura. Observando, las columnas son idénticas y, por tanto, son dos facturas que están duplicadas. Directamente las he eliminado del conjunto de datos.

También, a simple vista, se observa que hay valores nulos en las columnas store\_location y product\_category. Para rellenar los valores de store\_location tan solo he tenido que fijarme en store\_id, pues esta columna contiene tres valores (el 5 corresponde a Lower Manhattan, el 8 a Hell’s Kitchen y el 3 a Astoria), por tanto, al disponer del store\_id en aquellas facturas cuyo store\_location es nulo, se puede rellenar con facilidad. Por otro lado, para el product\_category ocurre igual, pero con product\_id. Cada valor de product\_id está asociado a un product\_category, product\_type y product\_detail. Por tanto, de nuevo, al disponer del product\_id en aquellas facturas cuyos valores son nulos en product\_category, es posible rellenar esta última columna con facilidad. Si el product\_id es 77 tan solo hace falta buscar otra factura con este mismo product\_id y rellenar con los mismos datos.

En cuanto al formato de las columnas, he ido una por una identificando su formato y cambiándola al mismo (por ejemplo, en el caso de unit\_price convirtiéndola a formato moneda, en el caso de transaction\_time a formato de hora, o en el caso de transaction\_date a formato de fecha, siendo así todo más legible. En esta última columna además había 5 registros desplazados a la izquierda, de forma que, primero, los he convertido a número para posteriormente cambiar su formato a fecha).

Otro aspecto para rectificar que se observa en un primer vistazo de los datos es el texto de la columna product\_category. En cuanto a la normalización, y concretamente a las mayúsculas y minúsculas, las categorías COFFEE, COFFEE BEANS y tea están escritas diferente al resto. El texto debe estar en minúscula con la primera letra de cada palabra en mayúscula, por tanto, es lo que he hecho, dejando: Coffee, Coffee Beans, Tea. Para la columna product\_type he realizado también varios reemplazos de forma que la primera letra de cada palabra en cada categoría este en mayúscula (por ejemplo, de Black tea a Black Tea/ o de Green beans a Green Beans). Para la columna de product\_detail lo mismo (por ejemplo, de Espresso shot a Espresso Shot).

Eliminación de columnas / creación de otras nuevas: La columna de store\_id la he eliminado directamente, pues me parece información redundante y para hacer este análisis en Excel con tener store\_location que realmente es la explicativa, es suficiente. Lo mismo me ocurre con product\_id. En caso de un análisis con SQL si pudiera interesar dejarlas, pero en este caso no. Para poder eliminar esta columna pero poder si quiero agrupar por el producto de café específico, lo que he hecho es crear una nueva columna llamada specific\_product, la cual va a ser el “significado” de ese código de la columna product\_id que identifica los productos, de forma que con este nuevo campo se va a poder hacer agrupación de productos específicos con un entendimiento mucho mayor, pues dispondremos del nombre completo del producto pudiendo identificarlo perfectamente, frente a disponer de un número que lo identifica pero que a simple vista no se conoce a qué producto pertenece. Para la creación de este nuevo campo lo que he hecho es concatenar los campos product\_category, product\_type, product\_detail (separados por comas).

También, he añadido una nueva columna: total\_price. Me interesa disponer del importe total de cada factura para posteriores cálculos. Esta columna ha sido calculada como el producto de unit\_price \* transaction\_qty.

Notificación anomalía observada: En el caso de la columna transaction\_id he podido identificar a simple vista que algo ha debido ocurrir con determinadas transacciones. De un primer vistazo, me pude dar cuenta de que el número de filas del Excel no coincidía con el ID de las facturas, pues al final del conjunto de datos había 149117 filas y el último ID era 149456, y por tanto, me di cuenta de que algo había ocurrido con ciertas facturas. El primer salto que observé fue 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.

Mi objetivo de análisis: Me interesa analizar a lo largo de estos 6 meses de estudio, en cual, de las 3 tiendas ubicadas en distintas localizaciones, se ha recaudado más y a qué se puede deber, tratando de que la que menos recaude logre aumentar sus ingresos.

Según mi contexto, creo que debo mantener las siguientes columnas:

* La de transaction\_id está clara, al final, es el identificador único de las facturas y es necesaria para poder hacer recuentos de transacciones.
* Store\_location: esta columna lógicamente también es esencial. Con ella se van a generar todas las tablas dinámicas ya que precisamente se está haciendo un estudio sobre los ingresos en las distintas localizaciones de las tiendas.
* Transaction\_date. En cuanto a esta columna, considero que es importante, pues podría ocurrir que una tienda venda más o menos en unos meses del año o en determinados días del mes (a principios, a finales, en los fines de semana…) y es interesante poder observarlo. En cuanto a esta columna, he decidido que puede ser más útil crear dos columnas nuevas: transaction\_day (1, 2, 3…) y transaction\_month (enero, febrero, marzo, abril, mayo, junio), eliminando transaction\_date. Me parece más útil ya que el año es innecesario pues todos los datos están recogidos en 2023, y además, así se puede realizar un análisis por separado (meses, días).
* Transaction\_time: esta columna también la considero relevante en el análisis. Es interesante observar a qué momentos del día las tiendas recaudan más o menos dinero. Por un lado, tener las horas desglosadas interesa ya que se pueden sacar valores como la moda interesantes. Pero, por otro lado, para lograr un análisis con mayor claridad lo que he hecho es crear una nueva columna: transaction\_time\_range que lo que hace es recoger el momento del día en tres rangos: mañana (de 6 a 12), tarde (de 12 a 18) y noche (de 18 a 21), pudiendo agrupar los datos en esos 3 intervalos sin tener demasiadas horas que observar, simplificándolo.
* Transaction\_qty: es interesante para observar si quizás una tienda es más propensa a vender en mayores o menores cantidades, y eso poder influir en sus ingresos también.
* Specific\_product: en cuanto a vender más o menos, es obvio que un factor muy determinante es los productos que vendas y la calidad de los mismos. Es interesante poder observar que producto específico es el más vendido. Quizás una tienda vende mucho más de un producto que otra, pues es su especialidad, y gana mucho con ello etc. De la misma forma de la que puede interesar analizarlo por productos específicos, también es interesante mantener las columnas de product\_category y product\_type, pues puede ser conveniente analizarlos con menor desglose, de forma más genérica. La columna product\_detail si se puede eliminar ya que disponiendo de la de specific\_product viene a ser básicamente lo mismo y es redundante tenerla.
* En cuanto a los precios, la columna de unit\_price puede ser eliminada pues ya dispongo de la de total\_price que es la realmente relevante para el análisis. La de total\_price es necesaria para hacer cálculos sobre los ingresos de cada tienda y, además, es interesante con ella hacer un análisis de estadísticos descriptivos. Cabe destacar que, aunque la de unit\_price se elimine, a partir de esta, he creado una columna nueva muy interesante y es el rango de los precios (por producto específico), ya que se puede con ello analizar si es que hay alguna tienda que sobretodo vende productos de precios elevados o al contrario, factor que claramente puede afectar sobre los ingresos.

En resumen, todas estas columnas son relevantes y pueden influir en que una tienda recaude más o menos dinero. La propia localización de la tienda puede resultar muy relevante a la hora de vender más o menos en determinados momentos del mes (si por ejemplo está en pleno centro de la ciudad venderá probablemente todos los días pero si está en un punto más turístico quizás vende más el fin de semana), en determinadas horas del día (si por ejemplo está en una zona de mucha industria con muchos trabajadores quizás vende mucho por la mañana desayunos, y menos por la noche), productos de mayor o menor precio (si es una zona de más dinero la gente se permitirá productos más caros…).

COMENTAR A LOS PROFESORES QUE HAY BASTANTE POCA DIFERENCIA ENTRE LOS INGRESOS DE LAS 3 TIENDAS (LA QUE MENOS ES LA DE MANHATTAN PERO POC TANPOCO MUCHO) Y SI ENTONCES TIENE SENTIDO QUE HAGA O NO EL ANALISIS DE ESO SI LA DIFERENCIA NO ES MUY NOTABLE. ADEMAS, SOLO HE DETECTADO QUE ESA PEQUEÑA DIFERENCIA PUEDE VENIR POR COSAS COMO QUE: LA DE MANHATTAN VENDE A PRIMERAS HORAS MAS QUE LAS OTRAS PERO A ULTIMAS HORAS MUCHO MENOS. TAMBIEN, QUE VENDE MAS QUE LAS OTRAS DEL PRODUCTO DE FLAVOURS (PODRIA HACER UN PREMIUM COMO TIENE LA DE HELLS Y PONER UN PRECIO MAS ALTO SOBRE ESE PRODUCTO SI GUSTA), Y QUE EN CUANTO A LAS CANTIDADES QUE VENDE EN CADA FACTURA TAMBIEN HAY UN POCO DE DIFERENCIA (DE 2 CANTIDADES VENDE MENOS PERO POR EJEMPLO 3 VENDE MAS QUE LAS OTRAS).

Por que cuando voy a insertar la tabla dinámica parece como que he seleccionado toda la tabla pero luego en la tabal dinámica no me aparecen las columnas que yo he calculado.

QUE HAGO CON LA COLUMNA DE UNIT PRICE SI HAY PRODUCTOS QUE SON LOS MISMOS Y APARECEN CON PRECIOS DIFERENTES. TENGO QUE CAMBIAR ESAS DIFERENCIAS, DEJARLO ASI O QUE

MAÑANA: SI AL FINAL HAGO ESE ANÁLISIS DECIR: CON QUE COLUMNAS HACER EL EDA NUMERICO (ESTADISTICOS DESCRIPTIVOS), CON CUALES HACER EL EDA CATEGORICO Y TEMPORAL (TABLAS DINAMICAS YA RELACIONADAS CON EL OBJETIVO) Y CON CUALES HACER LA MATRIZ DE CORRELACION

* Preguntar si las columnas que tengo están bien
* Preguntar sobre la de unit price que hago con ella (de ella va a depender la de total price y la del rango de precio)
* Preguntar si he hecho bien en los cálculos de estadísticas con esas 4 columnas y preguntar si es que no se puede calcular la moda en las categóricas
* Si la matriz de correlacion esta bien con estas 4 variables
* Preguntar si las tablas dinámicas están bien y tienen sentido las conclusiones. Y sobretodo, si puedo hacerla con la de la cantidad que es numérica, pero esque es un aspecto interesante a visualizar también entonces como lo hago si no.

**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 irrelevante, 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 “Observaciones” que me lo confirma. Ya en esta tabla dinámica puedo observar también que hay valores nulos en la columna de store\_location, 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 vacía, 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 “Observaciones” 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 “Observaciones” 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 dejo oculta.
* 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.

\*En esta base de datos concreta los nulos ha 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.