En mi caso, voy a usar los datos guardados en Oracle. Me he dado cuenta de que he tenido un problema de espacio por lo que solo se han guardado alrededor de 46000 filas de 51292 que había en el csv. Por simplicidad, se hará el *Data profiling & Eda* con las 46000 filas guardadas en la base de datos.

# Is each row an individual and complete shipping order?

**Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente**

Ilustración

Se utiliza la librería de *oracledb* para conectarse a la base de datos. Usamos la librería de pandas para crear un *dataframe*.

Tabla

Descripción generada automáticamente

Ilustración

Tabla

Descripción generada automáticamente

Ilustración

Si visualizamos la tabla vemos atributos como la fecha de envío y la fecha del pedido, coste del envío, tiempo de preparación, precio, … Por lo que se puede intuir que cada fila representa una orden desde que un cliente la pide hasta que es enviado.

# How many customers are in the data set?

**Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente**

Ilustración

Se hace una consulta en la base de datos para obtener **17415**.

# How many orders does each customer place on average?

*Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente*

Ilustración

Al hacer la consulta podemos ver que sobre unos 2.89 pedidos cada cliente de media.

# Are there columns with missing values? Which ones? Are they critical to the analysis we are doing?

***Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente***

Ilustración

******

Ilustración

Podemos hacer un reporte con la librería *ydata\_profiling*. Para hacer el reporte obtengo los valores de las dimensiones.

***Imagen que contiene Texto

Descripción generada automáticamente***

Ilustración

Vemos claramente que el 81.2% de códigos postales faltan. En el análisis que estamos haciendo no nos es relevante estos datos, podríamos obviarlo.

# Are there columns highly correlated? Which ones? If there are correlations, do they all make sense?

**Tabla

Descripción generada automáticamente**

Ilustración

Calculamos la matriz de correlación. En primer lugar, de forma trivial se puede decir que una columna tiene una correlación 100% consigo misma. Si consideramos una correlación alta a partir un 70% podemos ver que hay una correlación alta con:

* la fecha de pedido y fecha de envío
* el coste del envío y con el precio
* el id del tipo del pedido con el tiempo de preparación.

Realizamos un filtrado para verlo de forma más cómoda.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Ilustración

En primer lugar, tiene sentido que las fechas de pedido y de envío muestren correlación. Ya que lo normal es que la fecha de envío sea poco después de la fecha de pedido para todas las fechas de pedido

Por otra parte, también puede tener sentido que el precio de envío esté algo relacionado con el precio total, ya que muchas empresas ajustan el coste de envío al precio total de compra o productos más caros pueden ser más pesados o frágiles y al ofrecer un envío más seguro puede ser más caro.

La correlación entre el id del *shipment* y el tiempo de preparación también tiene sentido. Un envío *standart* de media debería ser mas lento que un envío en *first class*, por lo que es normal que exista cierta correlación.

After the initial data profiling task, we are going to carry out a brief **Exploratory Data Analysis (EDA) to understand the data at a deeper level.** Perform the necessary operations to answer the following questions:

# How much is the average shipping time of packages according to their ship mode?

***Texto

Descripción generada automáticamente con confianza media***

Ilustración

Obtenemos con una consulta las medias en días de tiempo de preparación, a simple vista parecen correctas. Como vemos si es el mismo día el tiempo es menor a 1 día. *First class* tarda alrededor de 2 días, *Second class* tarda alrededor de 3 días y *Standard class* pues tarda casi 5 días de media.

Vamos a comprobar que no exista ningún tiempo de preparación negativo.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Ilustración

Comprobando el reporte general que hemos hecho antes podemos ver que no hay valores negativos, pero si hay muchos valores que son 0. Podríamos retirarlos y volver a calcular la media. Sin embargo, como entendemos que el valor que se guarda es en días, puede tener sentido sobre todo para envíos de tipo *Same Day*.

# How much is the average profit that the company obtains per package ship mode? Is there any case where packages incur losses?

*Interfaz de usuario gráfica, Texto, Aplicación, Word

Descripción generada automáticamente*

Ilustración

Se obtienen con una consulta SQL la media para cada modo de envío y lo ordenamos de mayor a menor.

Interfaz de usuario gráfica, Aplicación, Teams

Descripción generada automáticamente

Ilustración

Como vemos el *profit* de medio de cada modo de envío es más o menos similar. Sin embargo, vamos a ver si existen valores anómalos para el *profit*.

Al ver los valores mínimos y máximos de *profit* podemos ver valores de hasta - 6599.9790 para el *profit* de envíos de clase s*tandart*. Es posible que estos valores hayan sido apuntados mal, o existan productos muy poco rentables y otros muy muy rentables. Para sacar una aproximación más real a la media podemos probar a retirar valores extremos. Filtramos para obtener los valores entre el percentil 5 y percentil 95:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ilustración

Las medias siguen siendo similares, lo cual tiene sentido, pero sale un vale inferior al de antes de media y valores mucho mas razonable en cuanto a *profit* se refiere para los productos.

Interfaz de usuario gráfica

Descripción generada automáticamente

Ilustración

Se observa que un número significativo de pedidos generan pérdidas. Según el reporte general, al menos 9,504 pedidos presentan un profit negativo, lo que podría deberse a errores en los datos, aunque no necesariamente implica que sean incorrectos.

Una posible explicación es que el costo de producción o preparación del envío supere el precio de venta final, lo que puede parecer un error, pero no siempre lo es. Muchas empresas adoptan estrategias en las que asumen pérdidas en ciertos productos para maximizar ganancias en otros.

Por ejemplo, una maquinilla de afeitar que se vende a un precio bajo, incluso generando pérdidas. Sin embargo, el verdadero beneficio viene después, con la venta recurrente de cuchillas de recambio, compensando con creces la pérdida inicial.

# How much profit do the 10 best clients (profit-wise) generate compared to the rest?

***Texto

Descripción generada automáticamente***

Ilustración

Con esta consulta puedo obtener el resultado de los 10 clientes que más *profit* han generado. El resultado obtenido es:

Tabla

Descripción generada automáticamente con confianza baja

Ilustración

Si quisiéramos compararlos con el resto de los clientes podríamos hacerlo de la siguiente forma:

Texto

Descripción generada automáticamente

Ilustración

Para hacer esta comparación hacemos lo siguiente:

1. Calculamos la suma del *profit* que generan los 10 mejores clientes
2. Calculamos la suma del resto de clientes
3. De esta forma podemos ver que los mejores clientes representando tan solo un 0.57% de todos los clientes generan un 3.16% de todo el *profit*. Lo cual es bastante

Podemos visualizar un resumen:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ilustración 20

# How much is the total profit per market?

***Interfaz de usuario gráfica, Texto

Descripción generada automáticamente***

Ilustración

Podemos ver con esta consulta la suma del profit para cada mercado. Si asumiéramos que todos los datos del profit están apuntados correctamente podríamos concluir que Europa es el mercado más rentable. En caso de que cosiderasemos que algunos profit son demasiado altos o bajo y sean probablemente un error de toma de datos podríamos hacer el mismo con estudio obviando valores extremos.

***Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente***

Ilustración

En este caso, no sería tan útil como en el cálculo de la media, pero si asumimos que los extremos son valores mal apuntados entonces nos puede servir, aunque sea para dar una idea de valores mas cercanos a los reales, aunque no valores reales. De cualquier manera, el mercado con mayor *profit* sigue siendo Europa.

Focus now on the market that provides the greatest benefit:

# How has profit evolved over time in this market?

Tabla

Descripción generada automáticamente

Ilustración

Podríamos visualizar el *profit* a lo largo del tiempo de esta forma. También podríamos crear un campo llamado *order\_date*, agrupar por meses y ver el *profit* a lo largo del tiempo. Usaremos *matplotlib* para visualizarlo de forma más cómoda:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ilustración

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración

*Profit* de Europa desde enero de 2012 hasta diciembre de 2015. Podemos ver que la tendencia es positiva y parece que hay picos por estación.

# What are the best-selling categories (by quantity) in the region of that market that has placed the most orders?

***Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente***

Ilustración

Con esta consulta obtengo que para Europa la categoría con mas ventas es la de material de oficina.

# Are the best-selling categories those that generate the highest total profits?

Como podemos ver en el *dataframe*, aunque haya aproximadamente el triple de pedidos para material de oficina los márgenes de *profit* deben de ser mucho menor de lo que se obtiene por producto tecnológico haciéndola más rentable.

# Based on all the answers you have obtained, what recommendations would you make to the managers of your company in order to improve results?

* En primer lugar, recomendaría a la empresa a empujar los esfuerzos hacía los aparatos tecnológicos sobre todo en Europa (Ilustración 26) que parece ser la categoría de productos que mas beneficios aporta el mercado más rentable.
* Se puede observar que los meses de junio y septiembre suelen ser malos para las ventas (no siempre, pero la mayoría) por lo que aflojaría todo el tema de la producción, marketing, … (Ilustración 25)
* Los meses más rentables parecen ser siempre uno o un par de mes previa a la caída de los pedidos en los meses de junio y septiembre (Ilustración 25).
* Los 6 primeros clientes que más *profit* generan son US central (Ilustración 18), sin embargo, son de los mercados que menos pedidos realizan (Ilustración 21) es posible que los compradores estén dispuestos a comprar más compulsivamente y gastar más o sean más fieles, por lo que se podría a lo mejor empujar el mercado hacia allí.
* No recomendaría aplicar muchos esfuerzos en la región africana ya que parece la menos rentable (Ilustración 21).