Assessment Aproach

**→ First steps and thoughts:**

Having a dataset in a csv and looking at the available technologies for the task i chose to go with Python

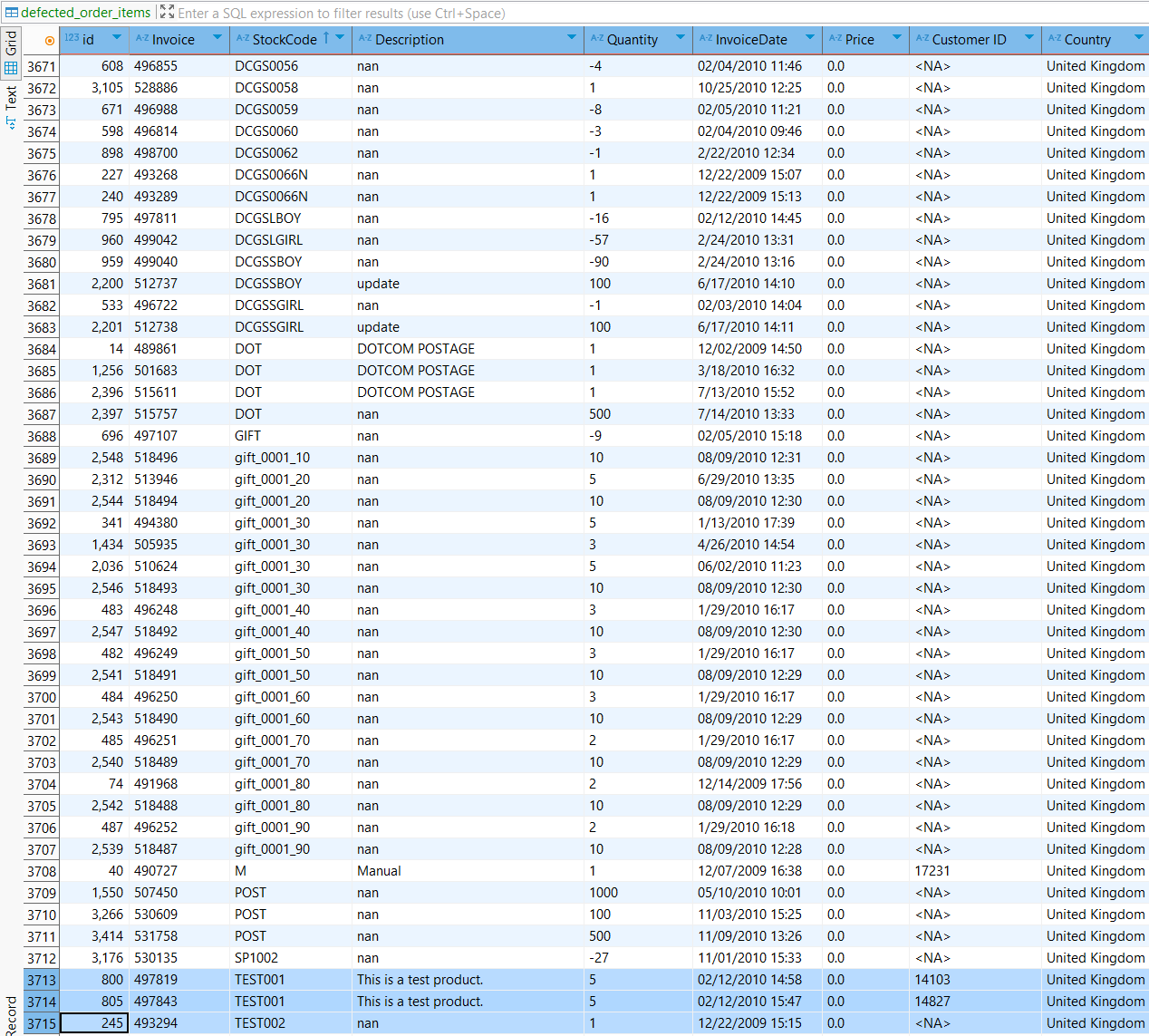
paired with DuckDB,pandas as they work great with csv/xlsx  
and i find the combination very convenient for exploring data fast.

**→ Assumptions:**

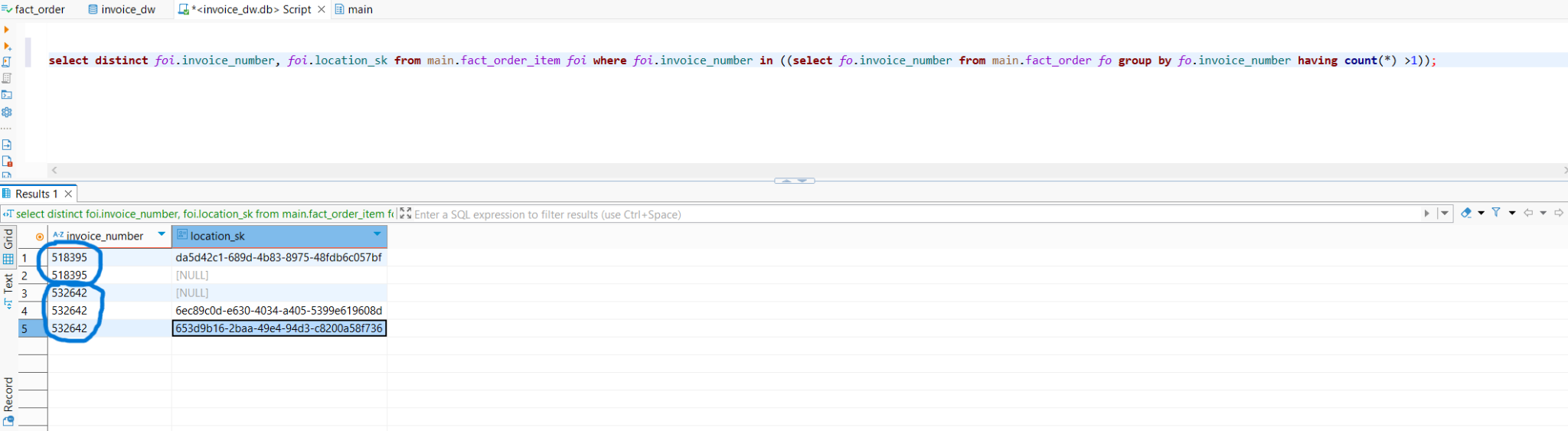
1. The data looks like a dataset from amazon’s retail department’s orders,  
   with a blend of both **line items (products)** but also with some entries that looked like **business actions & charges** (discounts,damaged good returns,post office fees, bank charges)
2. Seeing multiple **Invoice** (ids) made me consider grouping the items,charges,actions,etc.. to orders.
3. Since i dont know how they were exported (versions of programms etc., or if a human has tampered with the data before i got my hands on it i would expect even highly defective data

**→ Abnormalities & Analysis**

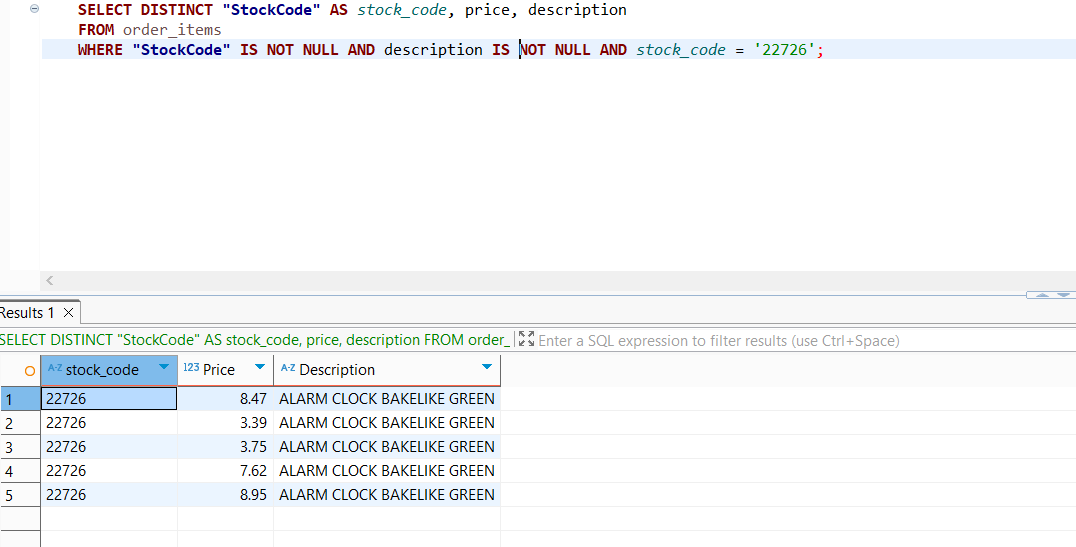
After looking at the data that did not pass the first filtering i found loads of products without either price,quantity,country,customerID and combinations of those.Here’s an example of these types of invoice which are NOT actions or charges and resemble normal products with “normal” alphanumeric stock code,quantity etc.



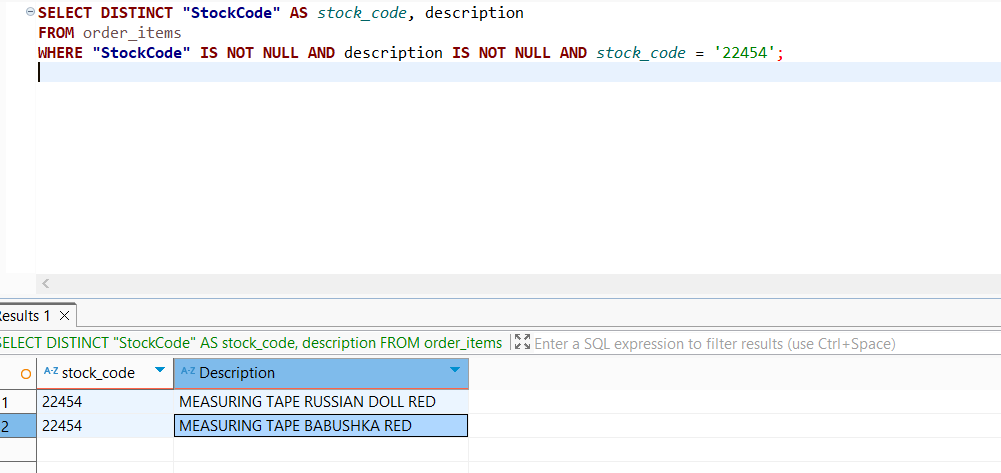
Further more i found multiple countries (which in some cases was the same country spelled differently or with other format like United Kingdom → U.K)



Also seen here having a country value null on a record would cause problematic aggregations (Revenue,orders by country)   
because we would have to opt out the nulls, causing missing line items from the actual orders total amount

Another interesting abnormality is products with multiple prices probably caused by changes on the price or applied discounts  
(insufficient business information and investigation to justify root cause)   


Same type of abnormality but with rather interesting solutions

apply to the same product having divergin-similar descriptions   
where in actuality its the same product.

Such issue could be resolved with normalization of the product either using the latest found description of such products and propagating it to its copies or by using an LLM that would eventually classify them as a single new product referring to these.  
  
**→ Conclusion**

With enough business requirements and time the defected data could be organised and cleaned a lot better but for the purpose of the assessment i chose to handle the breaking ones and focus on mentioning a few of the other findings to keep it short.  
A good example of non-breaking data existing in the set but rather questionable for reporting can be seen in one of the aggregations i made (TOP 5 selling products for each month of the first 4 months of 2010) where the DOTCOM POSTAGE fees are found in the top 5 in some months. Thus we cannot talk about “Product” aggregation explicitly