TASK 0

GETTING COMFORTABLE WITH THE DATA

Fabiana Caccavale
Matteo Gioia
Martina Manno
Vincenzo Junior Striano

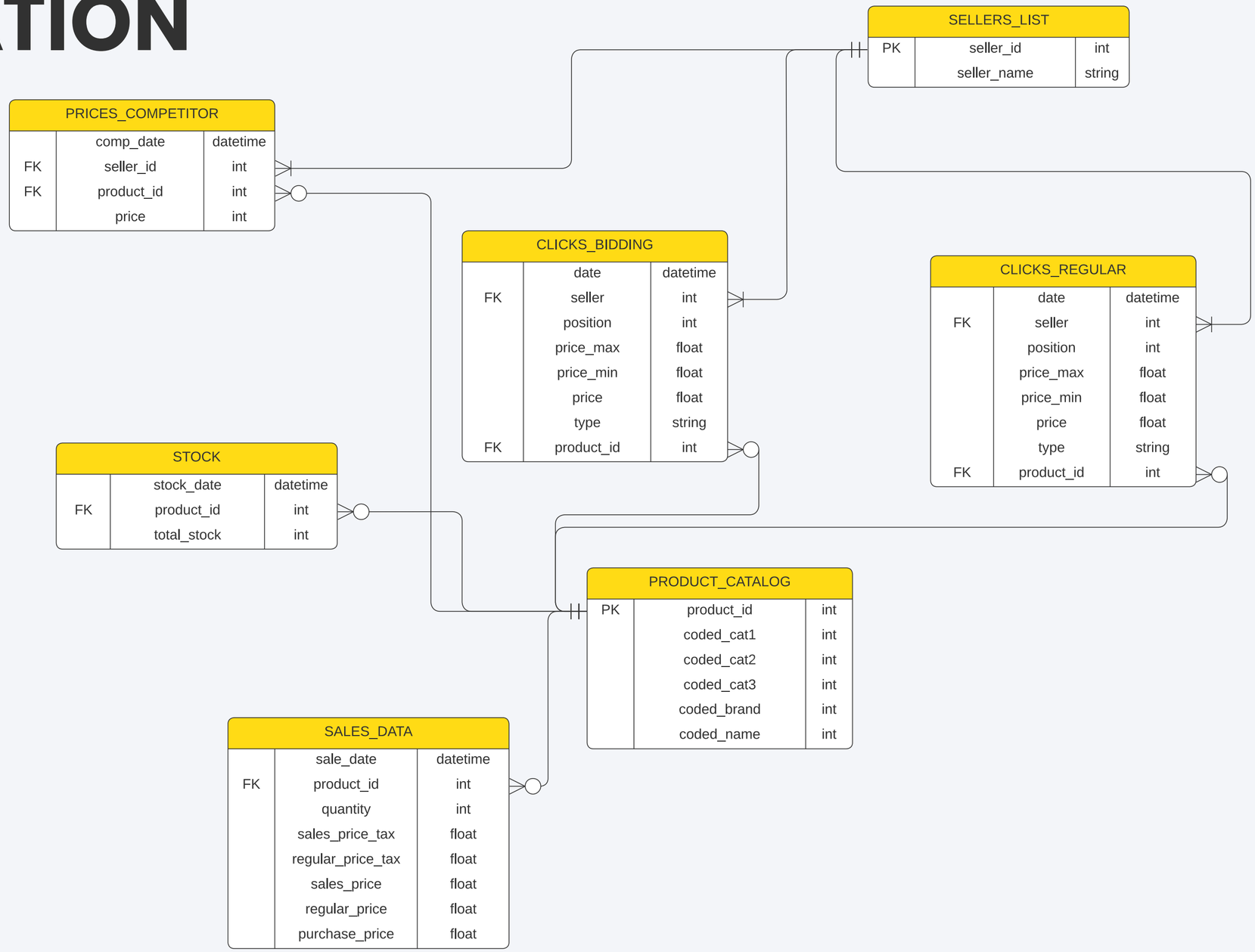
DATA FAMILIARIZATION

Are all the variables definitions clear?	All the variables have been understood
Sample vs All Datasets	We choose to work with the full datasets since we were not sure that the sample datasets were representative or balanced
Are the variables values consistent with what we expected?	Mostly yes but there were few cases of inconsistency (e.g. taxed prices = or < than untaxed prices; cases where products with stock < 0 are sold)
Do I need any kind of data transformation?	• Some columns needed data type changes for joining • In *clicks_regular* we transformed position by adding +3 because the bidding positions were 3 and in regular they started at 0. That allowed us to merge the *clicks* datasets
What is the time span of each dataset?	Time spans were different across the database (*clicks_regular* in the range [01-2021, 01-2022]; *clicks_bidding* [04-2021, 01-2021])
Findings	• Given the same product_id, the regular price changes periodically but it's shared across the sellers • In *price_competitors*, prices were reported for 6461 unique products. Of the remaining 1068 product_id's no info were provided • In *clicks* only 6485 unique product_id's were reported out of 7529
General Doubts	• "Meaning of Price Competitors"? * • Whom "sales_data" dataset belongs to? * *Question uploaded in the Learn.Luiss Forum)

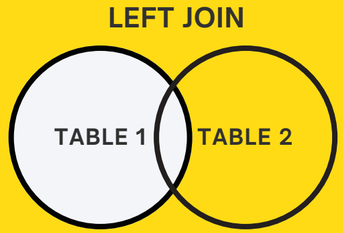
DATA INTEGRATION

The **ER model** shows what are the entities and the relationships between them.
It was the starting point to understand how the datasets could have been merged.
For example, in *product_catalog*, product_id is the **primary key**, or the unique identifier for a product.
In *sales_data*, product_id is a **foreign key**.
What are the relationships?

- How many unique product_id's have been sold? *Zero to many*
- How many unique product_id's are in a single *sales* observation? *One and only one*

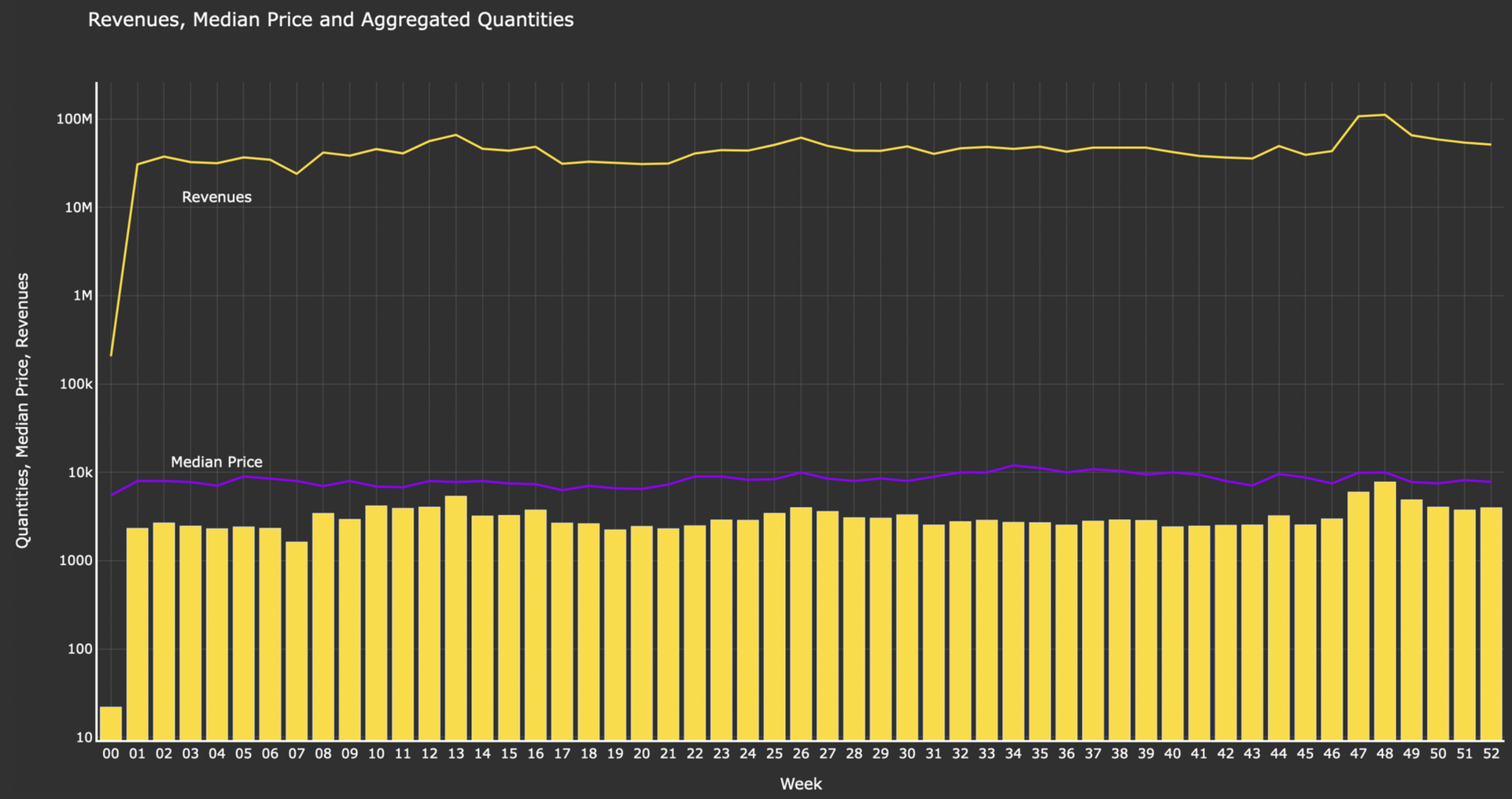


DATA INTEGRATION

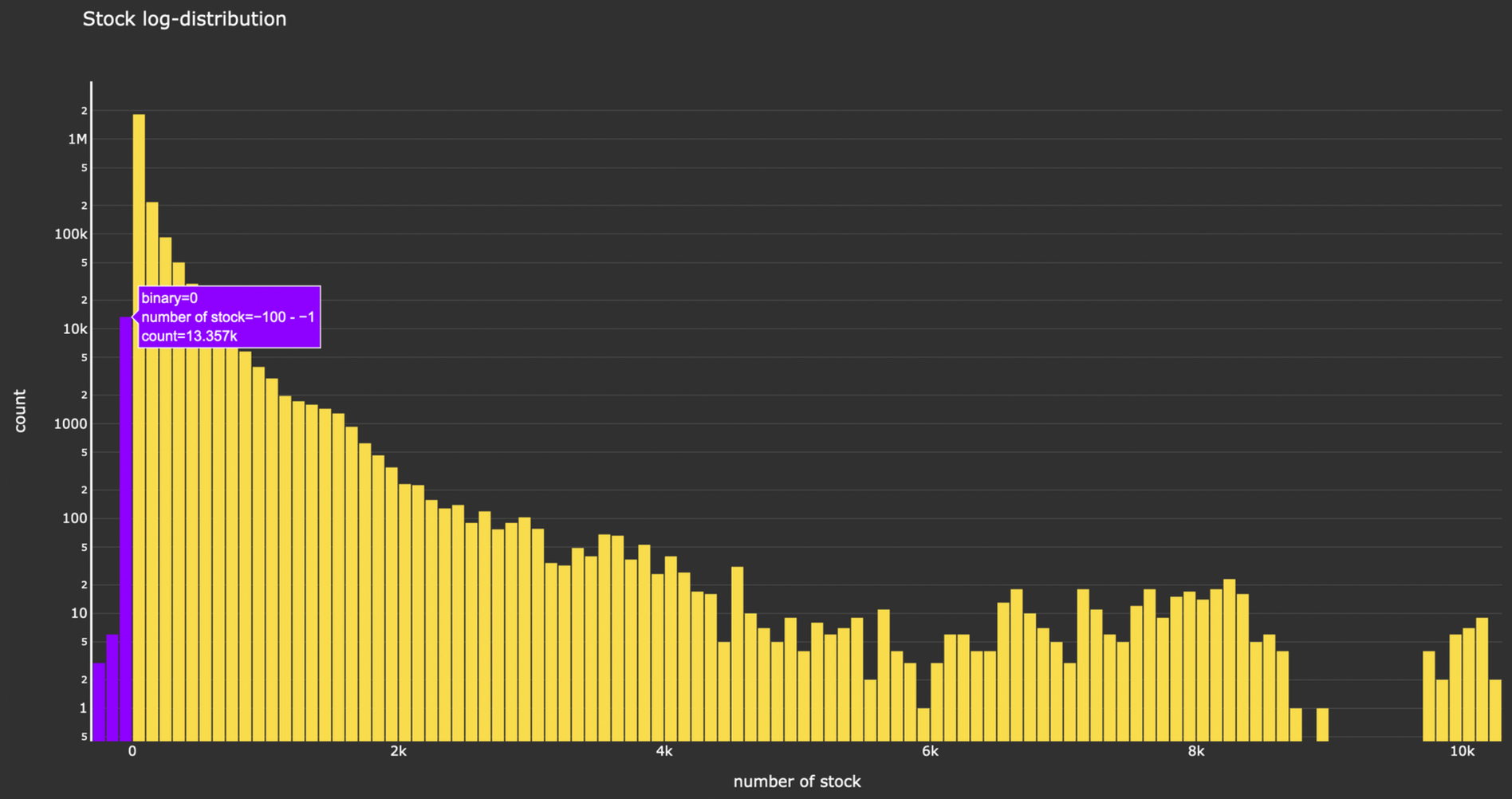
Which attributes allow to integrate the data?	product_id, date, and seller
Are all the data sources useful to integrate?	The only datasets which can be joined are the ones where product_id, date and seller are shown (e.g. *sales_data* does not contain seller and its records cannot be matched with the datasets about *clicks*)
What is the best way to merge the datasets with respect to my business needs?	Some data are split across multiple datasets. The best way to merge them is to use an outer join (typically a left join) used to match rows from different tables which share a key and retrieve data that are missing in one of them. For instance, in the *clicks_regular* there were 1093244 missing values for price. 87166 of them were retrieved by performing a left join with the *prices_competitor* basing on the match with date, seller, and product_id. 
Do I notice strange or duplicated values after the join?	By now, we haven't noticed any anomaly
Are there inconsistencies between the same attributes in different datasets?	The same attributes are consistent among the 7 datasets (e.g. all product_id's are present in the product catalog; all sellers are in the seller list etc.)
Can you add new attributes to datasets from existing ones?	Yes, it is possible to compute new variables by performing mathematical operations on existing attributes (e.g. revenue, taxation etc.)

DATA CLEANING

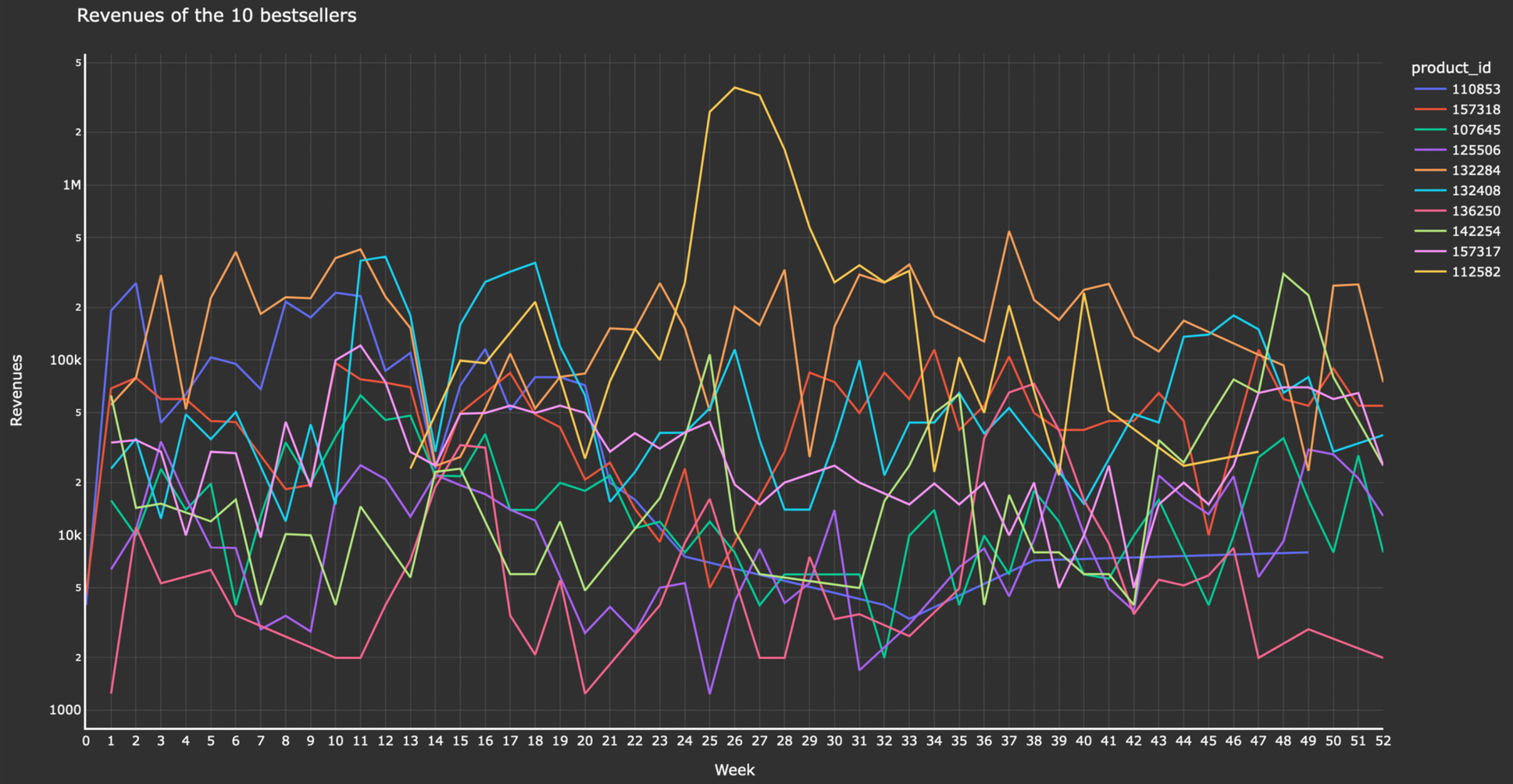
Describe the data	The datasets provide data on the online sales of an e-commerce firm specializing in electronic products. Moreover, a broader perspective of this market is available through price comparison website data
Do I notice extreme values?	An analysis of the *price_competitor* dataset to display the price range of each product according to the seller_id will be performed. By now, since the number of products is huge, we have decided to visualize the revenues/quantities for the 20 best-selling products (slide 8)
How did we approach null values?	Our approach is to keep as many observations as possible, even those showing missing or inconsistent data, which maybe be derived from other datasets to make further analysis. The datasets appear to be mostly complete, *clicks_regular* is the only dataset with missing values which we tried to fill in with some hypotheses. In *clicks_regular* there are also 124 observations where the price, price_min, price_max, and position are 0 which we considered unusable
How have we handled inconsistent values?	The values of the variables are mostly consistent, although there are some anomalies that we did not expect. Some examples below: For example, in *sales_data*, the minimum price of regular_price is lower than the minimum price of sales_price. It means that the minimum list price is lower than the minimum price after a promotion is applied. Moreover, the minimum value of taxation is 0, but this does not seem real because all goods are taxed. In *stock* the minimum value for the stock is negative so that the distribution appears right-skewed (slide 7).



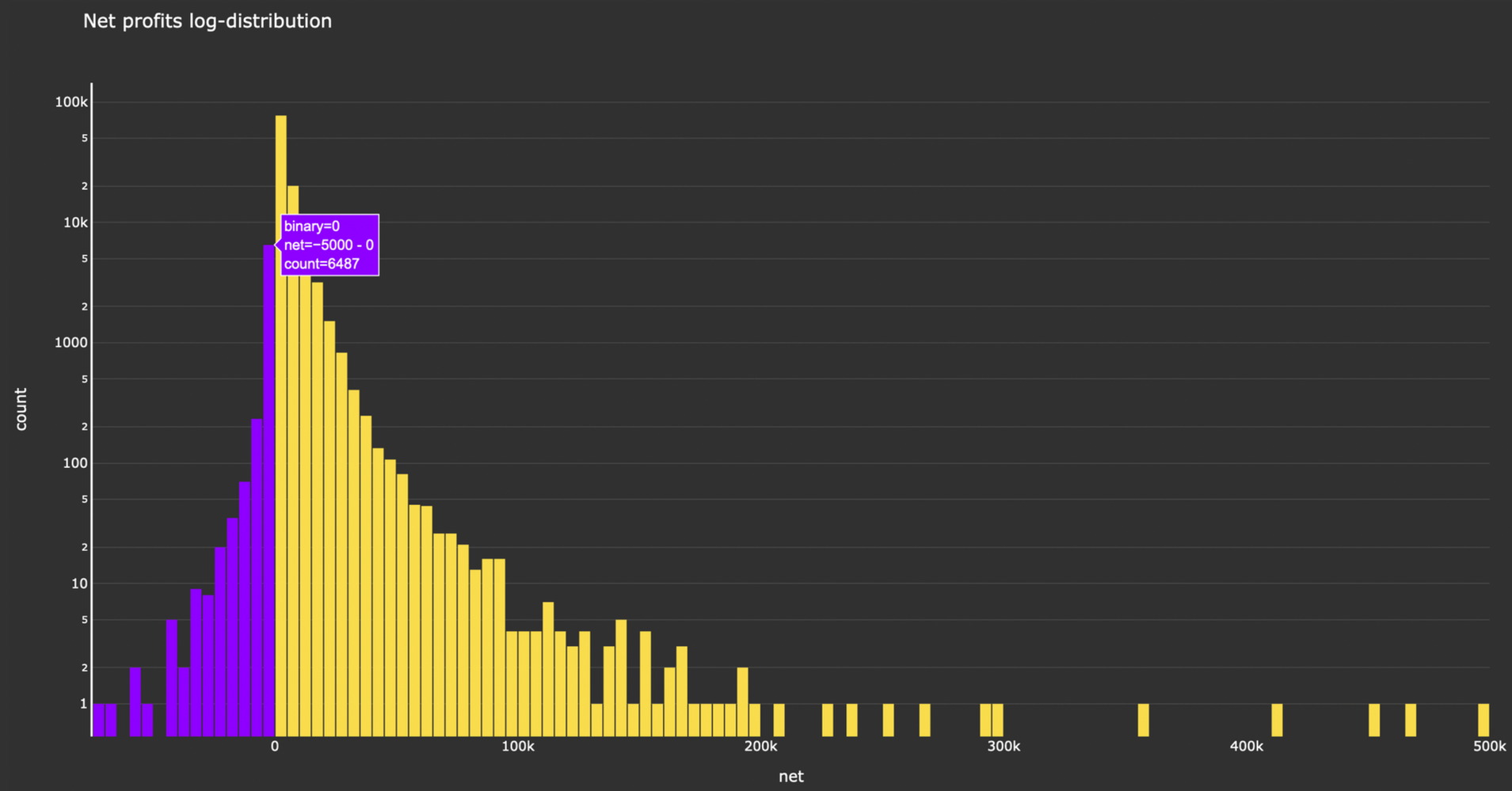
For each week are plotted the aggregated revenues, the median price and the quantity of sold products



Stocks log distribution barchart. In purple are shown negative, anomalous values



Weekly aggregated revenues from the 10 best sellers (quantity-related)



Earnings and loss log distribution barchart. In purple, the lossess

APPENDIX DOCUMENTATION

https://github.com/vincenzojrs/Alkemy-Spark-Project/blob/main/documentation/documentation_wip_task0.pdf