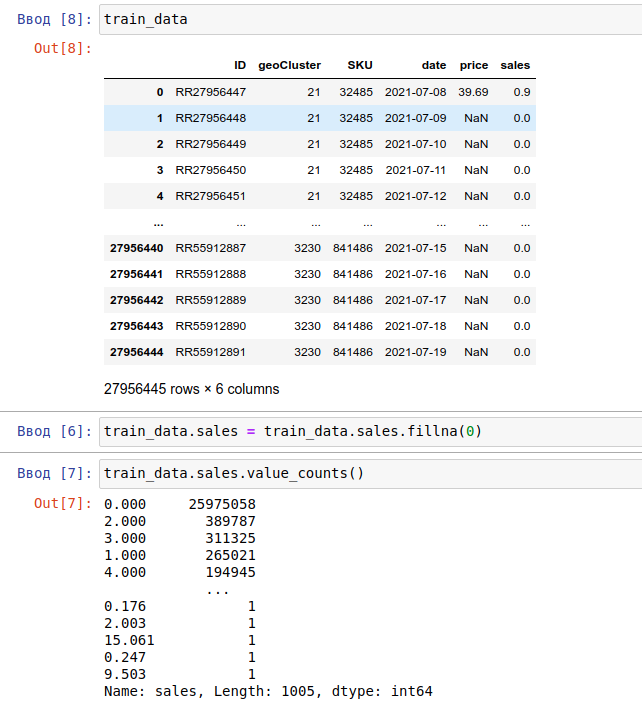
**Exploratory Data Analysis**

We began to work with the data with understanding, that from almost 28M data rows 26M rows contain 0 sales values.



One row represents a data for a ***specific day*** for specific cluster and specific SKU.

Also during some more deep analysis we found out, that many SKUs have one same problem. At the same geoCluster same SKU but after some sales have 0 sales and 0 prices.

There are 2 options why this behaviour can occur in the data:

1. Some issues with ETL and data preparation.
2. SKUs dont exist after some sales session in current geoClusters.

There is some example with banana.



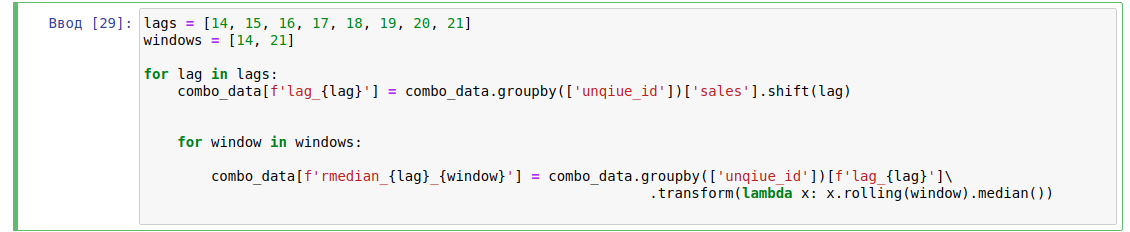
**Our Approach**

As a baseline we decided to use an LGBM model and provide some all features from train table and some features from SKU table, like 

but we didnt get any good performance.

So we decided to create lag features and consider only best 800 SKUs sales. The main idea is to create predictions for a specific day and specific SKU with specific cluster, but considering sales median for the past 14 and past 21 day + sales lag for the last 14 days.

In code it looks like this:



This features gave us the main baseline with 0.97 MAE on leader board.

But, the main problem is that we dont consider the less popular sales for SKUs. So as the next improvement we decided to use autoregression approach based on AutoArima for the least 200 SKUs.

As the result we received 0.906.

Finally, we gained 0.89 with blending auto arima results with the most popular 800 SKUs, that had the best sales and had zero sales less then 3 days only + ARIMA for the least sales SKUs.

**What didn’t work**

1. We tried to add more features from SKU table, but that didn’t work.
2. We tried FB Prophet model, but that didn’t help us.
3. One of the best ideas, but didn’t improve the metrics is to create models for different geoClusters and cities - however it is good for blending ideas.