# Transactions dataset research task

**We would like to test your general ability to manipulate data, perform research, build and support hypothesis based on data & data-driven story-telling skills.**

**In case you have any questions feel free to contact me directly by email: ivan.luchko@boosta.co**

In this task you should perform a small research based on the order peace of transactions dataset stored in **transactions_dataset.csv** file. Each record corresponds to separate transaction (order) characterized by its time and final status (is_canceled=1 for canceled orders). **number_of_paid_orders_before** describes customer experience of interaction with system before the current transaction. You have a piece of dataset, so for some customers records start from **number_of_paid_orders_before > 0**.

Please feel free to use numerical and visual approaches (plots, charts ect) to present your ideas and insights. Shortly discuss main dependencies, conclusions and insights received during your research. Also you are encouraged to provide some ideas which might explain those dependencies.

You can use any tool you prefer to perform the research Python, R, others, (Python is preferable). Please provide a short report and scripts/models you use.

1. Investigate **transactions time series**:

1.1. calculate average orders number placed per hour, day, week
1.2. plot placed orders time series (grouping data in bins by hour, day, week) – any ideas why dependence looks so?
1.3. investigate seasonality: daily(from day hour), weekly (from day of the week) - any hypothesis/conclusions/ideas which explain this behavior ?

2. Investigate **cancel_rate**: % of orders which were canceled

2.1. calculate cancel_rate for the whole dataset. What about **standard error, confidence intervals**?
2.2. investigate cancel_rate vs number_of_paid_orders_before (number_of_paid_orders_before altering in [0:10] range) – any ideas why dependency looks so? What about standard error?
2.3. compare cancel_rate for different months. – any ideas why dependency looks so? Any connections/correlations?

**TIP:** status *is_canceled* can be treated as a sequence of values which follows the *Binomial distribution*, where *p=cancel_rate*

3. Investigate **time between customer's placed orders – dT** (how often does the customer place orders)

3.1. Plot dT histogram, calculate dT mean & percentiles (time after which q={10, 20, … 90} % of customers have returned)
3.2. compare retention time period for **FCO** (1st customer's order) vs **not FCO** (2nd customer's order and more)
3.3. investigate what is the fraction of 'multi-orders' for which the time between two sequential orders < **dT_multi,** and it is short enough to consider two orders to be placed simultaneously. What **dT_multi** would would you choose and why?

**TIP**: some customers do not return after completing an order at all.

4. Investigate customer **retention_rate**: % of customers who return after completing an order

**TIP**: probably, **for each order** you would need to assign *customer return status* - **returned_after** similarly to is_canceled status
- for some order customer is considered '**returned_after**' if he placed one more order later within **dT_loss** time period. What customer lost time period **dT_loss** would you choose and why?
- some orders – last records of the dataset should be excluded from the retention rate calculation if not enough time has passed yet to consider customer lost if he is so (at least according to the dataset, where last record=current time)

4.1. investigate retention_rate vs number_of_paid_orders_before (order number for customer, altering in [0:5] range)
4.2. compare customer retention after order was canceled vs successfully finished. What about **statistical confidence?**

**5. Feel free to perform additional your own research in case you have any interesting hypothesis/insight you would like to prove/visualize using given dataset.**