**Press script**

Forms a table of programs with characteristics based on collected and processed data on publications in the press and social networks, and forms a press index based on them. Important! Only pre-premier data is processed, including the premiere.

The data folder contains four necessary files for work:

• press\_data - file with press data;

• FormattedPress\_Social - file with data on publications in social networks;

• LSD\_group.assignment - the file from which we need information about the groups to which particular series belong, depending on their position on the contribution curves;

• spends - stored here information about the date of the premiere.

Algorithm of actions:

1. Data preparation

• Load the necessary data.

• Combine files about social networks and the press.

• We get rid of data that do not contain data on Circulation, since this is the main measure of press data evaluation, and Run date, because then we cannot relate them to either the pre-prime or post-premature ones.

• Add information about the location of the series on the contribution curves. And also we add an additional column in which we number the received groups from 1 to 8 - from the highest to the lowest.

• Add a day of the week column to store information about the number of weeks between publication and premiere.

• We leave records in which the number of weeks varies from 0 to 25 weeks (getting rid of those cases where the publication was made a couple of years before the premiere (yes, there are such)).

2. Formation of metrics to assess the quality of press data.

• Calculate the percentage of positive tonality - the ratio of the amount of circulation with a positive tonality to the total circulation for each series.

• Calculate the amount of circulation for each series.

• Calculate the number of unique outlets for each series.

• Calculate Recency\_week3 - the ratio of circulation three weeks before the premiere to all circulation.

• Divide the data frame with serials by two - with new (first seasons) and old (not first seasons)

• In each received data frame, we group the series according to their location on the contribution curves and scale the resulting features from 1 to 10.

• Since the new TV shows are much smaller than the old ones, and the contribution curves are as much as 8, we are grouping some curves for the new TV shows.

• Curves 1-4 "5+ contribution groups"

• Curves 5-6 - "3-4 low groups"

• Curves 7-8 - "1-2 lowest groups"

• We form a press index, which is essentially a weighted sum of scaled features. The index ranges from 1 to 10 - from the worst to the best result. Ratios used:

• SUM\_CIRCULATION\_SCALED - 0.4

• UNIQUE\_OUTLETS\_SCALED - 0.2

• PER\_POSITIVE\_TONALITY\_SCALED - 0.1

• RECENCY\_W3\_SCALED - 0.3

• Last step: we sort by groups, and by press index within groups