# EDA:

## Observations

1. Target correlates well with numerical variables: "dcoilwtico" (oil price), "transactions", "onpromotions".
2. Sales patterns differ by category (“family”).
3. Essential goods have a more correct distribution of sales.
4. Information about stores and holidays has a lower correlation.
5. Time period under consideration: January 1, 2013 – August 17, 2017.
6. Granularity level is date + category + store. Observations are daily for all categories and stores, regardless of the presence of sales.
7. There is a lot of data: about 3 million observations. Each category accounts for about 90000 observations.
8. Oil prices have missing values. They correspond to weekends and some holidays.
9. Causes of data anomalies:
   1. Earthquake in April 16, 2016 – sales of certain categories increased in the first weeks after.
   2. January 1 – in many cases there are no sales due to closed stores.
   3. Christmas Eve – sales are on the rise for many categories.
   4. Sales grow in the first days of the month, after the payment of salaries.
   5. For some categories, sales start later than 2013, for some categories there are no sales in the first half of 2015.
10. Most of the sales refer to large cities, they also have the most type A stores – the leader in the number of sales.

## Conclusions

1. One of the first models to try is the linear regression with the numeric features: “dcoilwtico” (oil price), “transactions”, “onpromotion” and the categorical feature “family” to which one-hot encoding should be applied.
2. It makes sense to build a separate linear regression with the three numeric above for each category (“family”).
3. Because of the correct distribution of sales, it can be expected that a simple linear model would work better for products of the first category. For the rest, it may be worth building more complex models.
4. Information about holidays and stores can be added to a more complex model to improve its quality assessment.
5. The fact that we have a lot of data allows us to apply complex models to solve the problem.
6. It is worth cutting off the old periods in the data for some categories.
7. Anomalous time intervals can be thrown out (the bad thing is that the holes are formed in the time series) or can be smoothed using a moving average.