# EDA:

## Observations

1. Time period taken under consideration: January 1, 2015 – March 11, 2015.
2. Amount of data: 1197 samples (59 days). 691 of them refer to the “sweing” department, 506 – to the “finishing” department.
3. Granularity level: date + department + team. Observations are daily (except Fridays, Friday is a day off) for all departments, but not for all teams. There are days when there is observation for the team with “sweing” department, but not with the “finishing” department.
4. The target correlates poorly with all of the numeric features. There is very weak positive correlation (0.25) with “no\_of\_workers” and with “over\_time” (0.19). There is a very weak negative correlation with “idle\_men” (-0.15).
5. The distribution of the target differs across departments and teams .
6. The distribution of features (especially, “incentive”, “wip”, “no\_of\_workers”, “smv”) differs depending on the department – in the “finishing” department values of these features generally well below the corresponding values in the “sweing” department.
7. Information about the targeted productivity (“targeted productivity”) and about the number of style changes (“no\_of\_style\_change”) has the weakest correlation with the target.
8. Only “wip” (work in progress) feature has missing values. There are 506 such samples and all of them refer to the “finishing” department.
9. Causes of data anomalies:
   1. February 19 – last working day before “big” weekend (2 days, February 21 – national holiday). It is possible that this influenced on the poor productivity in “finishing” department.
   2. January 20 – anomaly in the “finishing” department – can be explained by the very small number of data relative to other days.
   3. January 22 – Thursday – the worst day (the same as Sunday), when the average value of the target is worse than other days. No anomalies were detected that day.
10. Since all anomalies in the data refer to the 3rd and 4th quarters of the month, the fact that the 3rd and 4th quarters are the worst in terms of the average number of positive class responses is probably misleading. Due to the rather small dataset, where only two full months are present, these anomalies make the 3rd and 4th quarters worse than the others, although in January there is no significant decrease in the number of positive target values ​​during these periods. Probably, the sign "day" in this sense gives more useful information.
11. Teams with the lowest “incentive” (bonuses) values (6, 7, 8) have the lowest productivity scores. However, the team with the highest “incentive” (9) is below the median productivity level. Teams 1, 2, 3, 4, 12 are the best in the productivity. All but 4th receive bonuses above the median level.
12. There is no seasonality in the data (in the distribution of the target).

## Conclusions

1. One of the first models to try is the logistic regression with the numerical features “smv”, “wip”, “over\_time”, “incentive”, “idle\_men”, “no\_of\_workers” and the categorical features   
   “day”, “department”, “team” to which one-hot-encoding should be applied.
2. It makes sense to build a separate logistic regression for each department (observation No.5), with the above features for the “sweing” department and without “wip”, “idle\_men” features for the “finishing” department (observation No.6).

When dividing the dataset by department, the correlation between some features and the target is more pronounced (for example, “incentive\_sweing” == 0.48).

Also, for the “finishing” department model, it is worth removing the “incentive” feature, because it is 0 throughout the entire observation period, except for one day (10 samples).

1. Other features: “quarter”, “targeted\_productivity”, “idle\_time”, “no\_of\_style\_change” can be added to the more complex model in the future to improve its quality (however, as it turned out, most of them do not have a special effect on the target, so this must be done carefully so as not to degrade the performance of the model).
2. It is worth trying to experiment with the “incentive” feature, because its correlation with the target on the entire dataset is very small for such a feature in terms of semantic load. As an option (EDA shows that it is good enough), add a binary feature (0 if the corresponding “incentive” value is 0; 1 if it is greater than zero).
3. Probably, instead of the one-hot-encoded feature “team”, it is worth trying to add a binary feature equal to 1 if the corresponding team is among the top-5 best, and 0 if it is not.