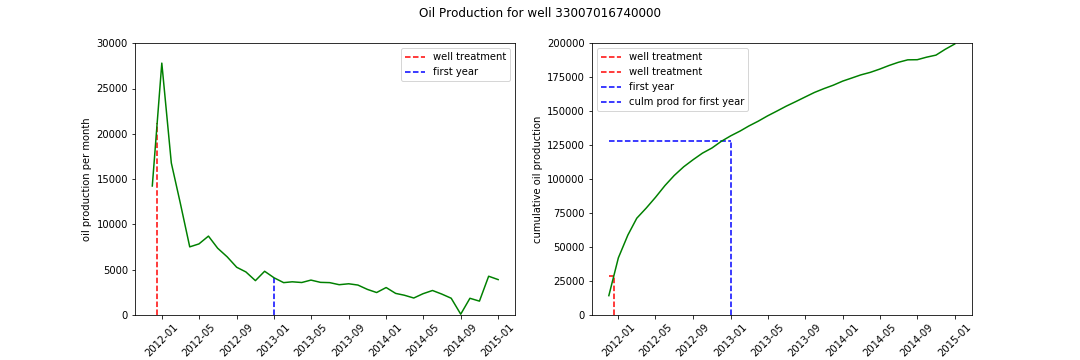
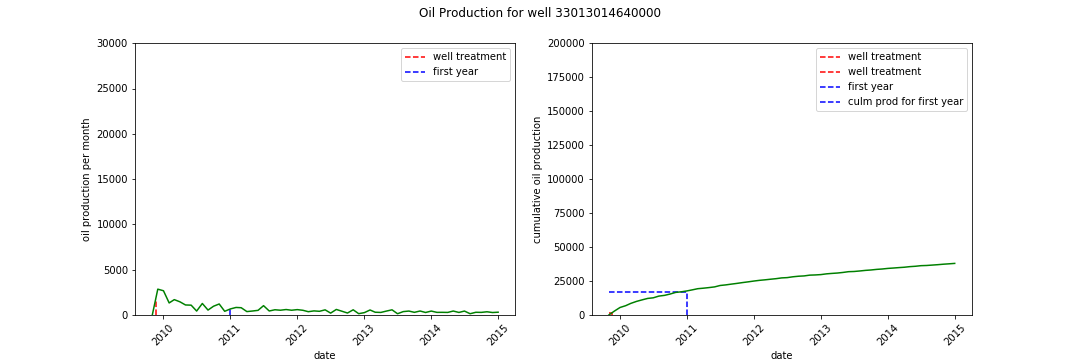
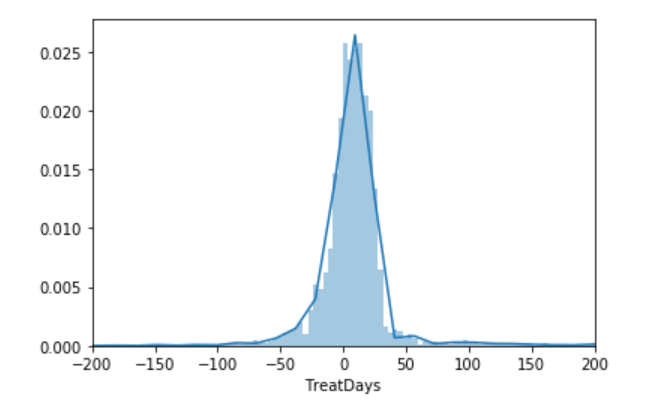
First, I need to visualize the oil production for different wells to get more familiar with the quantity I am trying to predict. The raw data is given as oil production for each month. I am simplifying the problem a little by predicting cumulative production for the first year. It is a good metric for overall well performance and can be used to separate good wells from bad ones.

The following plots shows the production profile for a good performing well. The left plot shows monthly oil production and the right plot shows the cumulative production over the same period. The blue line indicates the time stamp for the first year of production. The red line is the month of well treatment. 

Similar plot is created for a low performing well. 

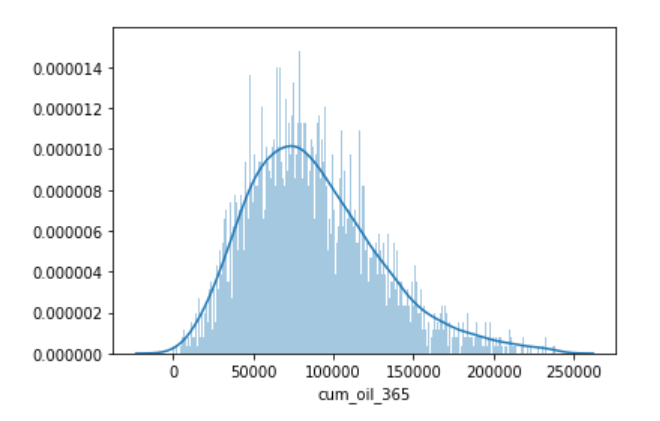
One interesting question to investigate based on these plots is theater or not well performance depends on how fast the well treatment was performed after the well was brough on production. The intuition tells us that the sooner well treatment is performed the better well production should be. I expect a positive correlation between number of days between first well production and well treatment and cumulation production for the first year.

First let’s plot a histogram of number of days between first well production and well treatment.



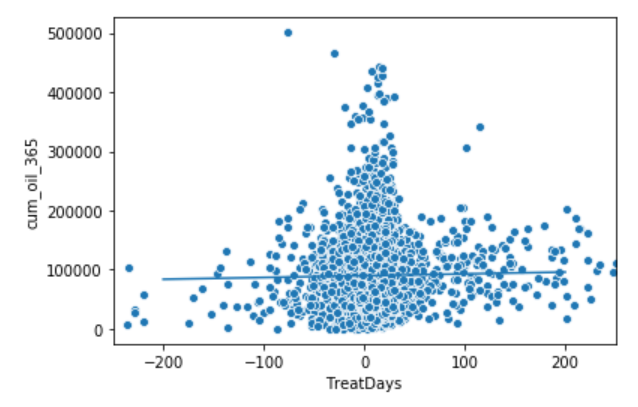
We can see that most wells have treatment job withing the first two months of production onset. It is interesting that almost half of the wells had treatment jobs preceding the production (negative values on the histogram).

The histogram of cumulative oil for the first year of production is well behaved distribution skewed towards the right-hand side.



We can investigate if there is really a significant dependency between the cumulative production of the well during the first year and how fast the well treatment was performed on the well.

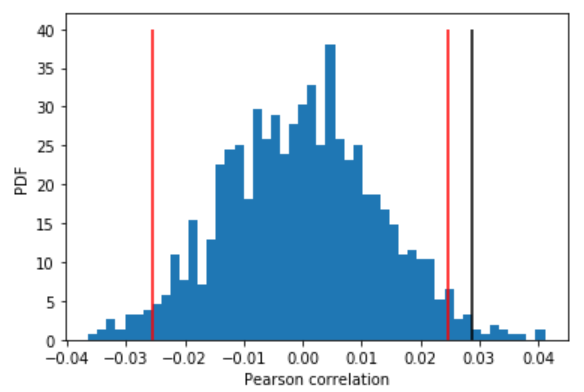
The cross plot of the two variables shows that they are not strongly correlated. Blue line show the linear regression. The pearson coefficient is 0.028 which indicates no correlation.



Our null hypothesis is that these two variable are not correlated. The observed correlation between TreatDays and cum\_oil\_365 may just be by chance.

To do the test, you need to simulate the data assuming the null hypothesis is true. To do so, we need to permute the cum\_oil\_365 but leave the TreatDays values fixed. This simulates the hypothesis that they are totally independent of each other. For each permutation, compute the Pearson correlation coefficient and assess how many of your permutation replicates have a Pearson correlation coefficient greater than the observed one.

The plot below shows the distribution of pearson coefficients for permutation replicas and an observed one from the data as a back vertical line. 2.5% and 97.5% confidence intervals are shown in red vertical lines.



The p-value is 0.012. It is quite small and suggests that our hypothesis is likely to be true. Cumulative oil production for the first year does not depend on when the treatment of the well happened.