Report on the MOCK data test

Prepared by: Pedro Monteiro

Date: 3 December 2015

Approach

I started by cleaning the data by removing all the rows with empty cells and plotting the time evolution of the total number of orders and drivers available. From this data I have noted that there are two regions with clear distinct rates: a region of no/lower growth, and another of higher growth (Fig. 1).

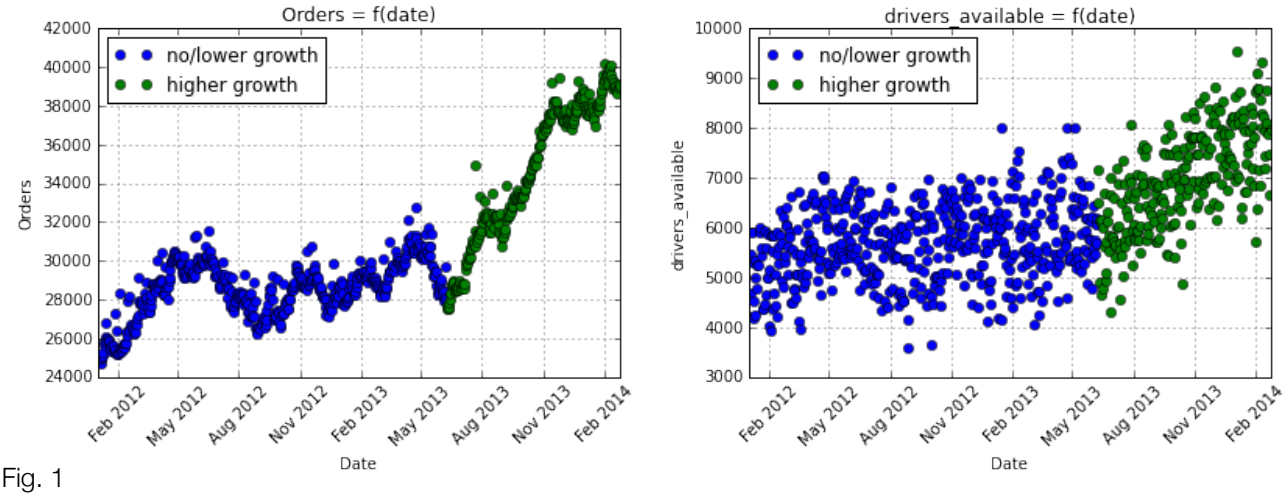

Fig. 1

Also, from the plot on the right, I noticed that there is a big spread of values of the number of available drives between consecutive days. This big variation in the number of drivers could potentially bring unpredictability to the business. There is no clear trend that a larger number of orders is accompanied by an increase in the number of available drivers to dispatch these.

To study this in more detail, I plotted the number of orders as a function of the number of available drivers (Fig. 2). This plot shows that, for a given number of orders, the number of drivers available presents a big variation.

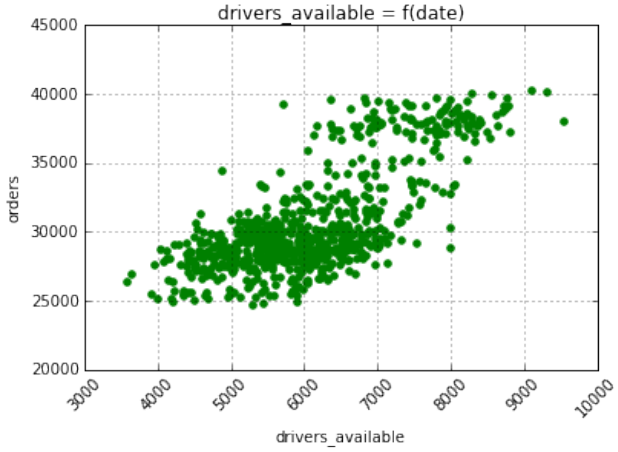

Fig. 2

Fig. 3 shows the mean values of the number of orders and drivers for different circumstances.

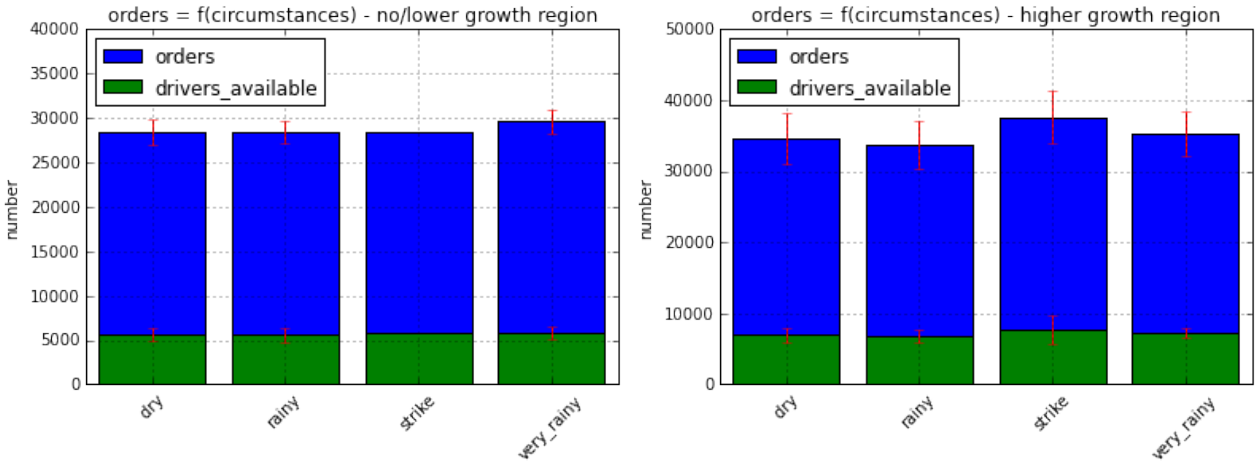

Fig. 3 The bars represent the standard deviation associated with the mean for each circumstance.

The event strike appears three times in the raw data: one appears in the no/low growth region, and two appear in the higher growth region. Therefore, there is not enough data to make statistical statements about this circumstance.

Regarding the other circumstances, the data shows that for the no/lower growth region, the dry and rainy circumstances have similar mean number of orders. On the other hand, very rainy days seem to have higher number of orders in both regions. All these observations suggest that the “strike” event does not refer to the company itself, but rather to an external body.

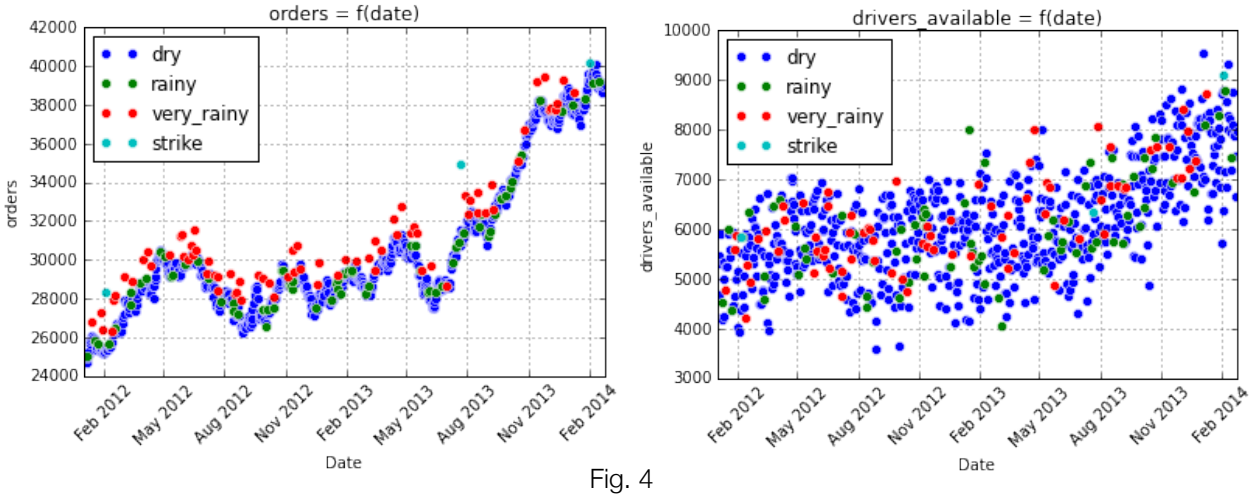

Fig. 4

The time variation of each individual circumstance as a function of orders and drivers available is depicted in Fig. 4. As it can be seen from the plots, the number of orders for the different circumstances present a similar trend, that is, a global increase in time. The data points associated with the number of drivers available (plot in the right) seem more scattered compared to the plot on the left. This is consistent with the behaviour seen in Fig. 2 and Fig. 1 right. The data from dry and rainy circumstances seem to overlap, whereas the data from the very rainy circumstance appears to be higher regarding the number of orders.

Statistical Analysis

	No/lower growth region date vs orders		Higher growth region data vs orders	
	dry or rainy	very or rainy	dry or rainy	very rainy
slope	4.523	5.122	42.1701	48.4863
intercept	-3294942.8966	-3734101.46732	-30966928.1562	-35609345.084
r^2	0.289	0.30015	0.893	0.9038
p value	1.352E-38	1.4956E-05	8.39182E-90	1.370E-10
standard error	0.3	1.1	1	3.7

To quantify the general trend that the data shows, I have performed a more detailed statistical analysis. The table above shows the linear fits used for each growth region. As mentioned before, the dry/rainy circumstance seems to overlap. Therefore, they were fitted together , while very rainy days were fitted separately. The fits show (please see ipython notebook) that the number of orders in the higher growth region is increasing at a higher rate than the drivers number, potentially hurting the value of the orders executed (there is a very high ratio of demand/offer).

How reliable is this trend? The linear fit for the number of orders in the higher growth region has a r^2 value close to 1, which means that the linear approximation seems particularly valid in this region. The general trend of the number of orders is therefore given by

orders = slope . date + orders(date=initial).

Summary

The transportation company in question is receiving more and more orders as time goes by, but the number of available drivers is not keeping up with this demand. Also the number of drivers varies a great deal regardless the circumstances. Whenever the circumstances are “very rainy”, there is the largest demand of orders. It would be useful to keep track of strike circumstances to make a more statistically valid statement in this case. However, the few data points available show an increase in orders as well.

To conclude, I would recommend to adjust the number of drivers to make the ratio demand/offer more efficient.