Supply Data Analysis

A modeling approach by Maurizio Ragusa

December 23, 2020

The key thing we need to do is to understand the dynamics between demand and supply. Here we'll make two statements:

1. The number of people who saw at least 1 car (denoted as S_1) is a function of how much waiting-for-booking drivers per hour there are in that time (denoted as W):

$$S_1 = F(W)$$

2. There is a relationship between the number of people who didn't see any car (S_0) and the incidence of booked drivers per hour. For instance, if the incidence of booked cars grows, we would expect that, with the same demand, the number of people who didn't see any car grows as well:

$$S_0 = G(B/O) = G\left(\frac{B}{B+W}\right)$$

where O is the total number of online drivers per hour and B the number of booked drivers per hour.

Understanding the nature of F and G is not trivial. Here we'll do some suggestions.

For (1), we suppose that the number of drivers seen by a customer follows a binomial distribution $B(W, p)$, where W is the number of waiting-for-booking drivers, and p is the probability that, for a fixed driver X, the customer can find X available nearby. We suppose that p does not depend on time, but it only depends on territorial factors, like the size of the city, the distance from which a driver X can be seen available for the customer, and the distribution of drivers in the city (which again, we suppose they don't change over the time). The probability p can be thus computed in the following way:

$$\frac{S_0}{S_1 + S_0} = (1-p)^W \longrightarrow q := 1 - p = \left(\frac{S_0}{S_1 + S_0}\right)^{1/W}$$

The value of q (see fig.1) is statistically quite stable: if we don't consider a limited number of values that are 0 (which depends on the fact that in those data we had 100% coverage, it maybe indicates oversupply), the median value is about 0.9 with standard deviation of 0.038. We can deduce that the value of q may be considered statistically meaningful.

As for (2), the relationship mentioned above is less clear and we could try to extrapolate it by data. Indeed, data suggest (Fig.2) there is a non-linear relationship between $\frac{B}{B+W}$ and S_0.

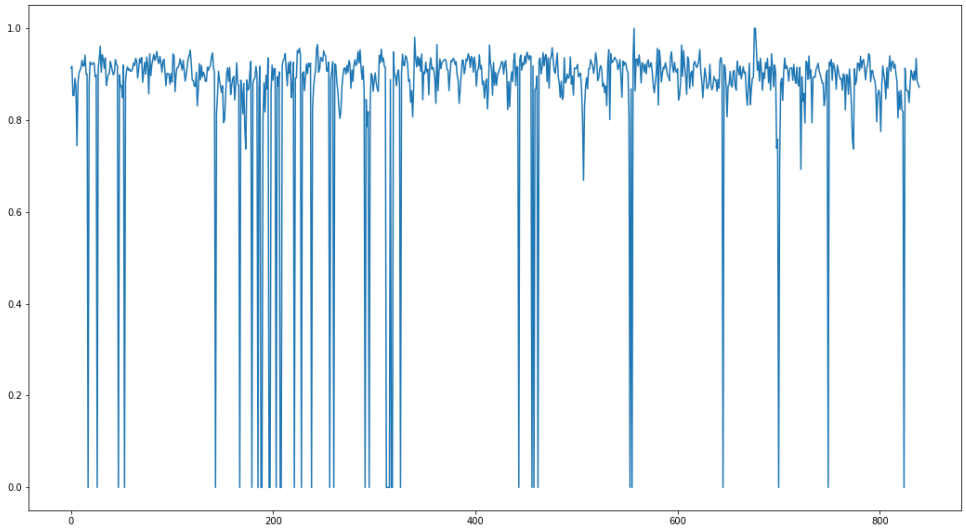

Figure 1: Given an online driver X, probability of not finding X nearby.

Operating a quadratic transformation on the x axis, we see that the relationship becomes linear (Fig.3). We then write:

$$\left(\frac{B}{B+W}\right)^2 = mS_0 + b$$

for some m and b which are the coefficient of the regression line in Fig.3.

Remark — *Note that this is a very brutal model. We suggest that much finer models can be found with a little bit of research, in order to better fit both empirical data and qualitative / theoretical considerations. This model can thus just be seen as an interesting direction one could explore in the future.*

1. Undersupplied hours

The most undersupplied hours are clearly those which highest not satisfied demand, that means customers who didn't see any free driver. In other words, hours when S_0 is highest. In the heatmap in Fig.4 we clearly see that the hours in a week with highest average demand are 8-9 and 18-19 from Monday to Friday (which is reasonable, since those are the hours where people move from home to work and vice versa).

2. 24-hour curve

The demand can be defined in an easy way, because it is the sum of both customers who have seen an available driver and those who have not, in other words, the demand D is:

$$D = S_0 + S_1$$

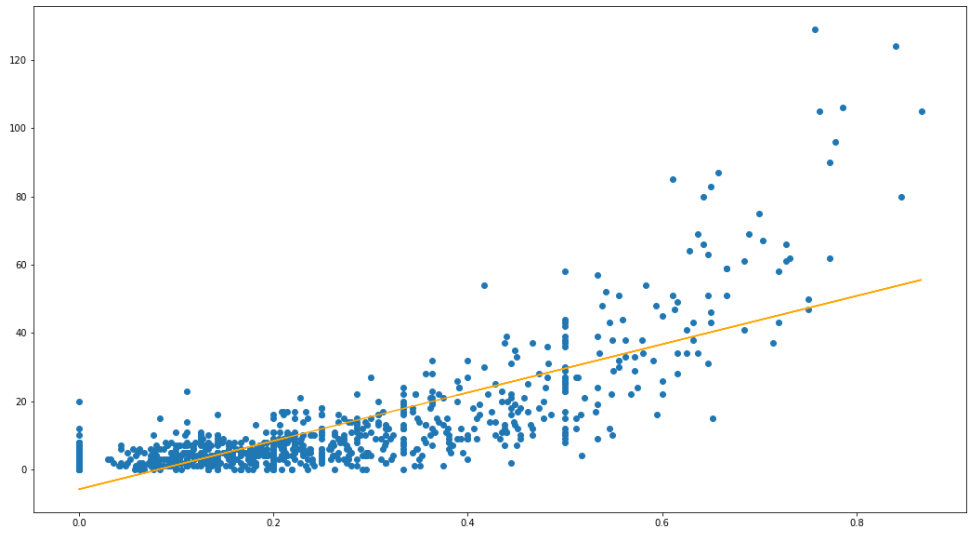

Figure 2: On the axis X, the waiting rate among all the online hours of drivers. On the axis Y, the number of people who saw 0 cars.

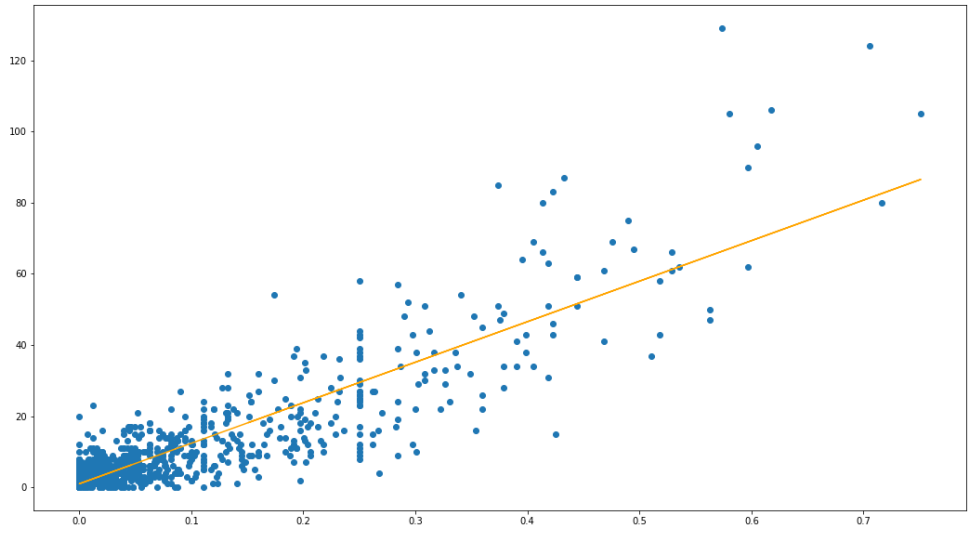

Figure 3: Same as before, but this time the axis X has been corrected by a quadratic transformation. The correlation coefficients are 0.87, which means quite high correlation.

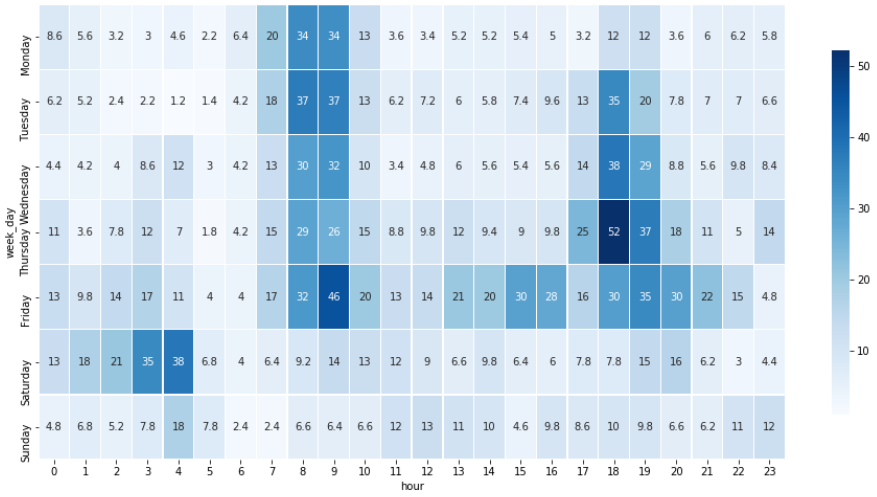

Figure 4: Heatmap of average people seeing 0 cars per hour/week

As for the supply, it can be considered as the number O of total online and available hours during the time period of 1 hour. Indeed, this may be seeing also as the average number of available drivers a customer can see in a random moment in that time period. The 24-hour curves (Fig.5), in which average data have been aggregated in hours, show that there is a spike in the demand at 8-9 and at 18-19, which is in line with what we found in the previous section. This spike in demand is not compensated by the supply curve.

3. Visualization of bad hours

We can use a simplified version of the previous table (Fig.6), in which there are less shades. Basically here there are just 4 shades of blue. The darkest blues are the hours who most need more drivers. In this way the heatmap can be easily understood by everyone in few seconds.

4. Keeping high coverage ratio

Now that we have a model, it is simple to compute the number of online hours we should increase to have an high coverage ratio during peak hours (say, the 36 hours with highest demand). First, let's fix a threshold τ for the coverage ratio. We fixed it at 80% but of course it can be changed. Since we know, as we noticed before, that coverage ratio depends on the total number W of waiting-for-booking drivers per hour, we can estimate the number W_{aug} of waiting-for-booking drivers needed to have the threshold chosen:

$$W_{aug} = \frac{\ln(1-\tau)}{\ln q}$$

Now, we can compute the number of online hours O_{aug} we need by inverting the previous quadratic relationship between the booking rate and the number of customer who didn't see a car (S_0):

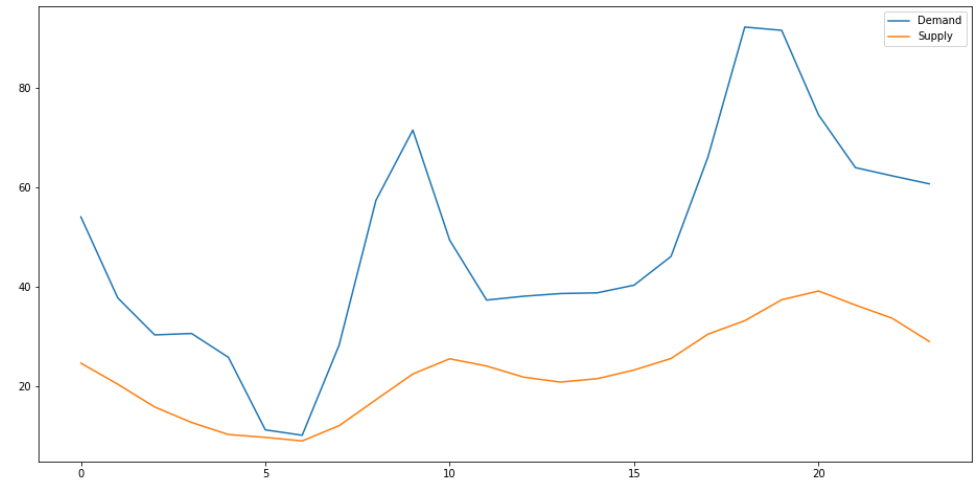

Figure 5: Demand vs Supply, on average, during a day.

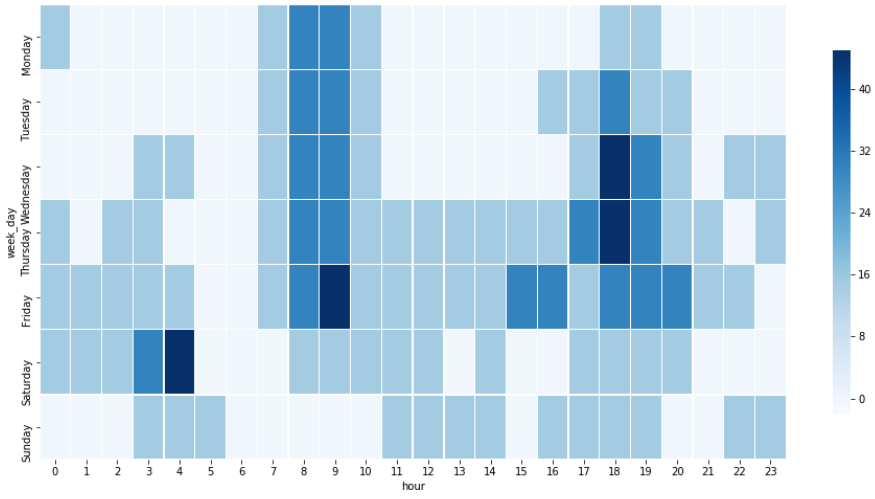

Figure 6: Simplified version of the previous heatmap.

$$O_{aug} = B_{aug} + W_{aug} = \frac{W_{aug}}{1 - \sqrt{mS_0 + b}}$$

Computation tells us that 420 hours should be needed to cover the 36 peak hours during the five weeks of data, which means, one average, 84 hours a week. For details about the computation, you can check the Jupyter file associated with this document.

Remark — *Note that, in the computations we made, the number B_{aug} of booking hours is essentially the same as the prior B. This is quite unlikely, which means that perhaps the entire model should be perfected before making any conclusions. As we said before, these ideas should just be taken as an interesting direction one could explore to better understand the nature of the problem.*

5. Giving money to riders

Let's assume that the finished ride per online hour RPH doesn't change if we change our supply. Let's also assume that the demand D remains the same. Under these conditions and knowing that our revenue is on average 2€ per ride, it's easy to compute our total revenue per online hour R:

$$R = O \cdot RPH \cdot 2$$

and now we can simply divide it by our $\Delta O = O_{aug} - O$ to get the amount of money we can give to drivers in order not to lose money.