# Report

Data preparation

To create histograms showing actual delivery length and prediction error I began with calculating time of delivery on every segment that is of type STOP and is connected to some order (has order_id).

Then I disregarded every record which duration of delivery was equal to or less than 0 seconds or was abstractly long as in 2-4 hours, considering them an error in data collection.

When recalculating seconds to minutes, time was rounded with formula where 2min30sec are rounded to 2 minutes, and 2min31sec are rounded to 3 minutes.

Result

The data manipulations described above left me with 2101 records. Histograms representing delivery length and prediction error are presented below.
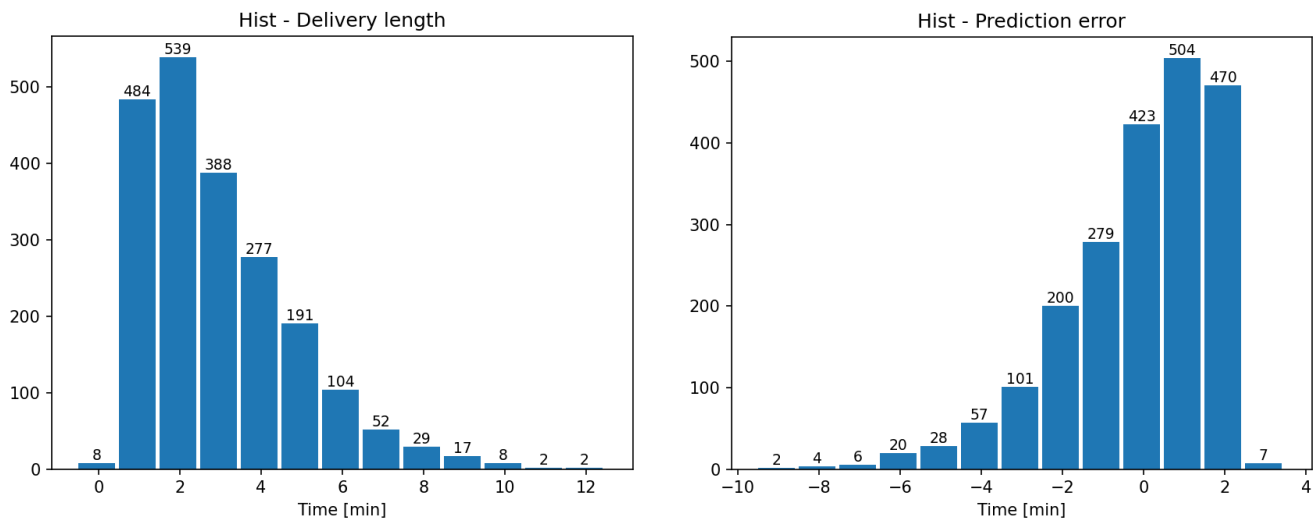
Analyzing just the above charts we can see that almost half of orders (49%) are delivered in up to 2 minutes (2.5 min. to be exact) and about 67% of all orders are delivered in up to 3 minutes (3.5 min.). There are 8 orders that appear to be delivered in under 30 seconds. We can consider them an error or assume that the driver made an excellent job in delivering order to single-family house.

The prediction error mostly appears in favor as for 67% of orders drivers had up to 3 additional minutes to finish the delivery. On the other hand there is a room for improvement, because for 33% of orders drivers needed from 1 to 9 minutes more than predicted to finish delivery.

Continuing to the next question, where drivers informed about deliveries taking longer in one of the sectors, I prepared a chart showing an average delivery time and how many orders where placed in each sector.

Deliveries by sector
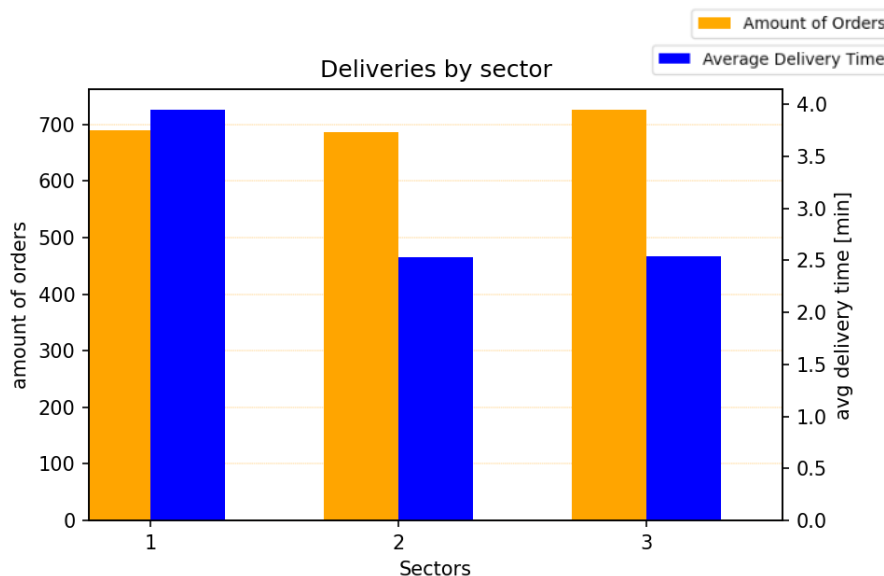
We can see there is about the same amount of orders placed in every sector so we have an even amount of samples from each area, witch makes calculated average time more believable. It shows that sector 1 has in average 1.5 minutes longer delivery duration.

I decided to look closer and check out each driver for each section (chart below).
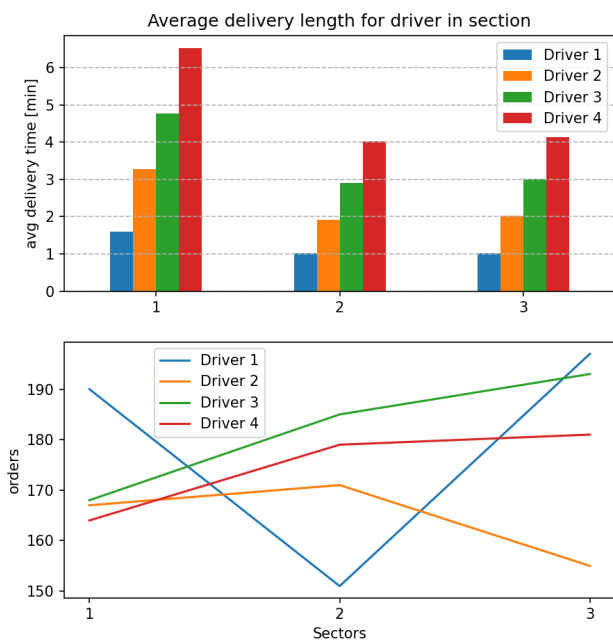
These chars give much more information. First, they support the conclusion that the 'sector 1' is the one where deliveries take longer. Second information is that is goes for every driver, so it's not the case of one driver inflating the over all average.

What I would do is I would sit with 'driver 1' and 'driver 4' and ask them about their delivery routine, because for some reason, 'driver 1' has the best outcome in each sector and average prediction error for his deliveries is above 0, so he almost always has time to spare. In contrast to 'driver 4' who's deliveries take the longest in each sector (4 times longer than 'driver 1') and is almost always behind the schedule (negative prediction error).

Meanwhile I would change formula for calculating "planned delivery duration" by including in the formula not only sector in which the delivery will take place, but also a driver that will do the job. This should improve the accuracy of prediction for each order.

Also to improve the average delivery duration in 'sector 1', system could be favoring assigning deliveries in this area to 'driver 1' instead of 'driver 4' (what perhaps is already done, assuming from the lower-left chart, orders delivered by drivers per each sector) and the 'driver 4' would relive 'driver 1' from as many deliveries in 'sector 3', where he has better outcome.

More information

Trying to find an explanation for why 'driver 4' is the slowest I decided to check weight of the orders they deliver. We can see his packages are a bit heavier in average than other drivers, except for 'sector 2' where 'driver 1' has heavier packages and yet is still faster then 'driver 4'. Although 'driver 1' delivers to 'sector 2' 15% less orders than 'driver 4'.
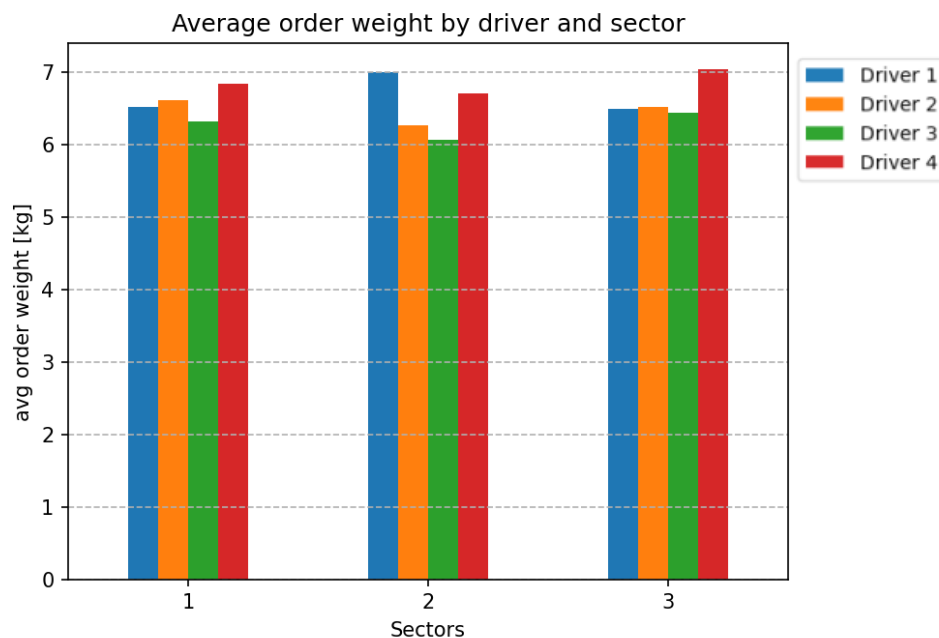
I also decided to check how much time they spend driving. Here I also disregarded segments where time was amounting to hours. It takes about 2-12 minutes to get through a segment. All the drivers spend similar time driving. 'driver 1' and 'driver 3' spent driving 3-6 hours more than the others. They also had to go through 30-40 more segments each.

Maybe there is a way of improving driving routes so that they won't spend additional time driving around.
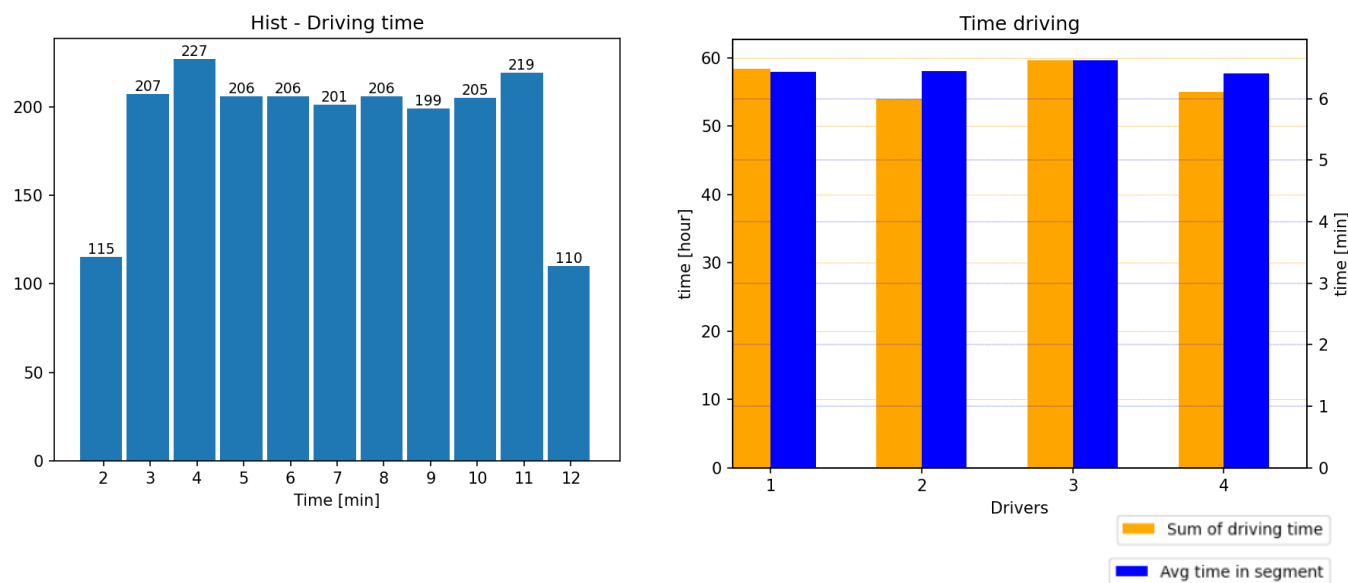
I also looked at the segments with type STOP in which there is no order delivered. I figured it either means that drivers got stuck in the traffic or made some additional stops.
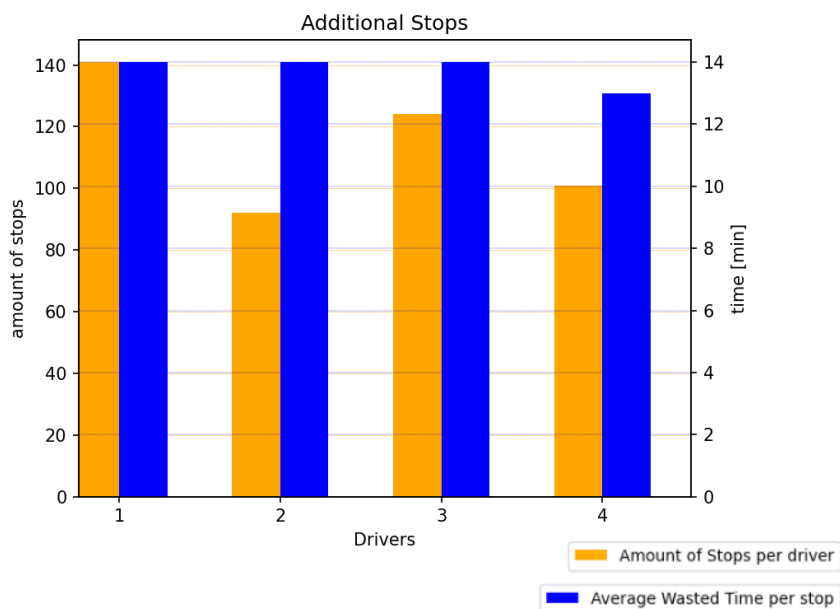
We can see that each driver stops in average for the same period of time. 'Driver 1' has the most additional stops, so perhaps he needs to change his driving route (same conclusion as from previous chart → some changes might reduce his working hours and speed up deliveries). 140 stops a month (February in this case) mean about 5 stops per day. Each stop last around 14 min. so it calculates into **over an hour of lost time a day.**

Additional info

If you want to see full SQL queries used for data extracting see file: queriesVisiualization.ipynb