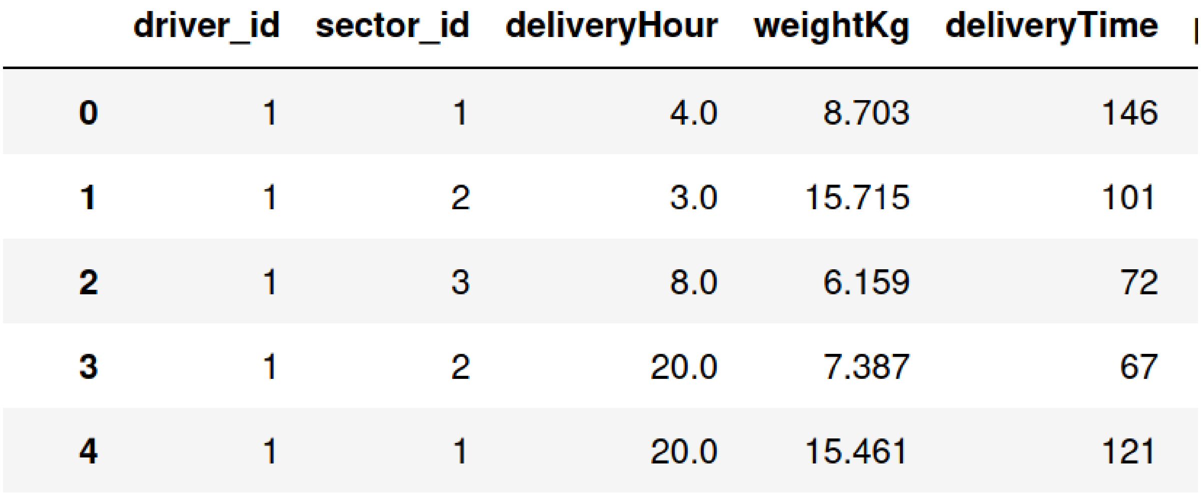
Building and verifying the hypothesis

1. Predicting delivery time per sector could be achieved by calculating the mean delivery time from all deliveries in each sector. Then when a new prediction would be needed the planned delivery would equal to the mean of previous deliveries in the sector where new delivery was placed.
2. A very good algorithm for predicting delivery times is a machine learning algorithm called Random Forest Regressor. In order to use it I had to create a new dataset using available data, as available raw data contains a lot of features that would not be useful for this algorithm. New dataset contained: driver id, sector id, hour when delivery ended and weight of delivery in kilograms (sum of product weight multiplied by quantity of the product in each order) of deliveries whose segment type was 'STOP' and where actual delivery time was between 0 and 30 minutes. I calculated delivery time by calculating the difference between the date when delivery ended and date when delivery started. After creating the dataset I have splitted it into two parts: training dataset (80% of whole dataset) and testing dataset (20% of whole dataset). Then the model was trained on training dataset.



	driver_id	sector_id	deliveryHour	weightKg	deliveryTime
0	1	1	4.0	8.703	146
1	1	2	3.0	15.715	101
2	1	3	8.0	6.159	72
3	1	2	20.0	7.387	67
4	1	1	20.0	15.461	121

First 5 samples from new dataset

Predicting delivery time for new orders can be achieved by providing data to the model, after that the model calculates results using parameters chosen during the training process.
Methodology for validating the algorithm: I verified the algorithm by calculating mean squared error (MSE). It is a very popular metric that is equal to the sum of all squared differences between value predicted by algorithm and actual value. Smaller MSE means a better prediction.
Random forest regressor MSE was 2600. In comparison, the MSE of planned delivery in the orders table is 13520. In conclusion Random forest regressor is over 5 times better than simply calculating planned delivery time based on mean delivery time of all previous deliveries.
3. Some deliveries take more because of numerous reasons. Here are some of them:
   - Some deliveries might be bigger and heavier than others.
   - Delivering during rush hours might take more time.
   - Delivering in crowded areas such as the city center might take longer as it can be harder for drivers to find a parking spot.
4. Ideas for additional data that could be collected:

- Information about deliveries during several parts of the day. For example there is little to none data about deliveries at these hours: 10am, 11am, 12am, 1pm, 10pm, 11pm, 12pm.
- Data about deliveries during other months could be helpful. Now the whole dataset contains samples collected in february.
- Information if the delivery target is an apartment building or a house.
- Experience and age of the driver.

5. Under-estimating delivery times might lead to delays in deliveries and to annoyance of the clients. Overestimating delivery times can result in elongated planned delivery times, which could lead to less attractiveness among competition and loss of customers.