# Part 3. Building and verifying the hypothesis

### 1. The current prediction algorithm is very naive. It calculates the mean from all collected data and applies it to every future order. We need to explore alternative ideas. One of them is predicting delivery times per sector. Describe how you would validate this hypothesis using available data.

Using available data, I'd validate this hypothesis by comparing the mean absolute error of these 2 methods which is the average of the absolute differences between predicted and actual values, showing how much predictions typically deviate from the real outcomes. So firstly, I'd calculate global, and sector means then compute the mean absolute error using scikit-learn python library and compare the results to if the hypothesis is true.

### 2. Using the data, propose some alternative method/algorithm that will predict delivery times more accurately. Describe the methodology to validate the new algorithm.

I would propose using a machine learning regression model, such as the Random Forest Regressor that predicts the actual delivery duration based on data features. We can extract a range of attributes from the data to serve as input features for the model. To validate that method, I would use a train-test split and calculate the evaluation metrics to interpret the results.

### 3. Why could some deliveries take more time? For example, some buildings don't have elevators etc. Describe your ideas.

I can think of few categories:

- Building related delays:
  - No parking nearby - may have to walk from a distant parking spot.
  - Unclear apartment numbers
  - Broken intercoms - Forces you to call or wait for the customer.
  - Security-controlled entry
- Traffic and navigation issues:
  - Traffic jams
  - Poor GPS accuracy
  - Road closures - Construction or emergency services rerouting the traffic.
  - Road accidents
- Customer-related delays:
  - Customer not answering the door or phone
  - Special request of instructions from customers – may take some more time than usual deliveries.
- Weather conditions:
  - Heavy rain – slows driving and walking.
  - Snow
  - Fog – reduces visibility.

### 4. What additional data would be worth collecting for future analysis of this domain?

It would be valuable to collect additional data like building attributes (if it has an elevator, floor number, type of building, security gates), order details (if it's fragile, refrigerated, huge in size). Driver experience level and vehicle type would also potentially have a big impact on delivery times. Other than that, external conditions like weather data and customer specific information (historical delivery time, preferences and their feedback on past deliveries).

### 5. What is the risk of over- or under-estimating the delivery times?

Risks associated with overestimating delivery times include drivers sitting around in between deliveries and less appealing delivery time options, which may cause customers to select faster alternative services.
Underestimating delivery timeframes carries the following risks: drivers rushing increasing the likelihood of accidents, poor customer experience (missing delivery windows, for instance), complaints, refunds and bad reviews harming the brand.