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.

Predicting delivery times based on sectors is likely to be more effective than using a global average because it considers local differences. Some sectors may have more complex infrastructure, which can increase delivery time. To validate the hypothesis, I would:

- select only rows where the segment type is STOP and delivery time is lower than 240 minutes/ not null/ not negative (to exclude potential outliers or incorrect records)
- group the data by sector_id – creating 3 sets
- divide each set to : training (80% of the data of the data) and test (20%)
- for each sector, calculate the average delivery duration using the training data
- predict the delivery time in the test set using this average
- use error metrics to compare predicted vs. actual delivery time and real data averages (R2, MAE and RMAE) to see how well the model can predict values
- compare the results with doing the same calculations without grouping the data by sectors

Using performance metrics helps to assess whether the model learns from the data or just makes random guesses. For example, if the R2 score is significantly higher for the sector-based model than for the global model, it means that sector-specific predictions are more accurate. This suggests that sector is an important feature in delivery time prediction and should be included in future models.

RMSE, on the other hand, emphasizes larger errors - in this case, it helps identify which method results in bigger mistakes in delivery time predictions.

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

ML model – using linear regression, Random Forest or XGBoost - these models can learn from historical data and consider different factors that may affect delivery time.

For example, we can use features like start time (time of the day), sector ID, driver ID (some drivers might work faster), and package weight (heavier packages might take longer). The model can then learn from these patterns and predict delivery time more accurately.

After training the model, I would evaluate how well it works using metrics like R2, MAE, or RMSE. This approach would give us a smarter, more personalized prediction model that adapts to different situations instead of treating all deliveries the same.

3. Why could some deliveries take more time?

- **heavier or larger packages** – drivers may need more time to handle and carry heavy or bulky items, such as large boxes or packages containing fragile goods like glass
- **the building is farther than usual from the parking spot** – if the driver cannot park close to the entrance, they need to walk a longer distance, which increases the total delivery time
- **winter weather conditions** - snow and ice during winter months can slow down deliveries. Drivers walk more carefully to avoid slipping and need more time to carry packages safely
- **waiting for a customer** - if the customer is slow to answer the door or if there's a conversation involved, the delivery takes longer than expected
- **searching for the right package in the truck** - when the truck is not well organized or carries many packages, finding the correct one can take additional time
- **gated or secured building** - entering properties that require access codes, intercom calls, or security checks

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

- **distance between stops (or delivery addresses)** – measuring the actual driving distance would help understand how long it takes to move between deliveries within a given sector and time of day. This could significantly improve delivery time predictions.
- **traffic conditions** - these can be used to improve predictions, as higher traffic levels may cause drivers to take longer than usual to complete deliveries
- **driver experience** – less experienced drivers may take longer to complete certain tasks compared to those who have been working longer
- **floor delivery** – whether there is an elevator or not, deliveries to higher floors usually take longer

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

- **under-estimating**
  assigning too many packages to drivers - tight schedule, stress and postponing the delivery to another day (not building trust with customers with promised delivery – lower customer satisfaction). Overall, it creates logistical challenges and disrupts route planning

- **over-estimating**
  assigning too few packages to drivers - lower productivity, increase in cost. It can also lead to slower delivery of all orders, especially when the system plans inefficient routes based on incorrect assumptions