Questions:

* Approach:
  + Option 1:
    - A row is one train number on 1 day, for example: Train 741 on 1/1/2021.
    - The dependent variable is a binary whether the train is on time or not (delayed less than 360 seconds)
    - This approach has the advantage that we can use train specific variables (length, departure time, capacity ...). However we feel the dependent variable is not representing our target well.
  + Option 2:
    - A row is all trains on a line during a peak hour period (morning/evening) in a day, for example: 14 trains in the morning rush hours from Ghent to Brussels on 1/1/2021
    - The dependent variable is the proportion of trains that were on time.
      * Correct target?
    - In this approach we can have the amount of P-trains a variable, which represents our target well. However we lose train-specific variables.
    - What kind of model could work for option 2?
      * Neural network? Boosting-family?
  + Alternative or extra: simulation and/or network model
* Covid:
  + Combine before and after? (paradigm shift)
* Ideal goal: specific P-train on specific hour and line -> what is the impact on whole network
  + Feasible at all?
  + How would we capture effect of train on other lines?
* Outliers:
  + Our outliers:
    - Q2 +- 3 \* IQR as limit
  + Jelena’s suggestion:
    - Use log transform, remove any that are past the line
  + Need to check validity of extreme outliers + explain
* Programming: errors for large datasets
  + When to switch to spark?
* Modeling feature importance:
  + Feature importance don’t contain p-values
    - How to evaluate P-train
  + How to check statistical significance of P-train impact?
    - Challenging because observations will always have some amount of P-trains
* General Project-Thesis admin:
  + Currently only working on project
    - When to switch? Timeline wise
  + What are the expectations in terms of presentation? (business & technical)
  + Expected length of thesis in pages

TODO for ourselves:

* Do trains change lines? 50A and 31 for a single train