PYTHON GROUP ASSIGNMENT

INFRABEL / NMBS

2022-2023

# Practicalities

## Form groups

* Form groups of **4 or 5 people**. Subscribe as a team to a group on Ufora (Groups > Group Assignment) **before the 12th of October at 12:00 am**. Students who have not enrolled by then will be automatically and randomly assigned to a group.

## Deadline

* The deadline of the group assignment is the **8th of November (12:59 pm)**. By then, you should upload your PowerPoint presentation on Ufora (Ufora tools > Assignments > Group Assignment) and are not allowed to make any changes to your code on GitHub (see later) anymore.
* The presentations will be held online and will probably take place on the **9th (and 10th) of November**. The exact schedule will follow later.

## Presentation

* At the end of the assignment each group will present their findings within 20 min.
  + **15 min presentation**
  + **5 min Q&A**
* Create a **professional** presentation
  + Presentation should be directed at a non-technical audience, however, questions might be technical as well.
  + Present as if you would present to the NMBS management. Try to interpret your findings and communicate a message that is relevant to them.
  + Make it visually appealing
* There will be a **strong** **time pressure**:
  + Knowledge to do the analyses is learned during classes
  + Limited presentation time, so it is necessary to select the **most relevant results**, without losing information
    - E.g. Often different plots can be combined into a single plot, saving space and presentation time and making comparison easier
    - Having backup slides with additional information might be a good idea…
* Do not limit yourselves to the obvious outcomes. If you know additional analyses that might be relevant and that are not mentioned in the goals below, feel free to do it. Moreover, go beyond simply reporting results, interpret them as well and search for links.
* Try to come with a **clear message** during your presentation.
* You will have to create your solution using a **GitHub repository**. GitHub is an online platform to facilitate coding projects which is very popular in the IT/Data Science industry. You will have an on campus introductory course on how to use GitHub. In the meantime, you are of course allowed to work outside GitHub. Once you create your repository on github.ugent.be, you should add Professor Van den Poel (dvdpoel) and Dylan (dvmulder) to the repository. The final solution should be included in the repository by the end of the project.

# Assignment

## Data preparation

1. Upload all provided datasets.

## Analysis

1. Plot the average delay time per provider (EUROSTAR, NMBS, THALYS).
2. Plot the number of stations per city for cities with multiple stations.
3. Plot the number of stations per city in function of number of inhabitants.
4. Plot the distribution of delay times.
5. Plot how the average delay time evolves through the week
6. Plot how this delay time evolves across the day for the five most used stations in Belgium.
7. Plot all Belgian stations on a map of Belgium.
8. Plot all train routes on a map.
9. Are some train routes more likely to experience delays than others? Visualize and proof statistically.
10. Plot how the number of stations differs across Belgian provinces using a color-coded heat map. Do the same while adjusting for the number of inhabitants.
11. Which cities are the worst with regard to access to train facilities? You can do this by calculating the travel distance, travel time, … Would you recommend based on this, and the visualization in Q5, to create some new routes?
12. Cluster the stations based on facilities. What type of stations exist? Are some stations ‘odd’ in the facilities they provide? What would you suggest to do with these? Is there a correlation between station type and number of daily trains?
13. Determine unique facilities that are very rare in orders of prevalence. Infer what may cause these facilities to be in place on their current locations. Are there possible other stations that could benefit from these facilities?
14. Plot the average delay time in function of the number of train rides per hour. Does there seem to be a correlation?
15. Is delay determined by possible delay in the previous station? (Hint: this is a form of autocorrelation).
16. Regress the number of facilities to both the number of daily trains and number of daily travelers. Do this using two univariate regressions and determine which covariate is the primary driver for number of facilities, based on the adequate goodness-of-fit measure.
17. Extreme delays heavily impact our business model. Visualize how these are distributed, both per incident type and geographical location. Are there any incident types / locations that we should pay special attention to? What actions would you recommend?
18. Last year we conducted a survey per station on how satisfied the average traveler is. We observed some large differences to exist, but are puzzled on what causes this. Proof statistically which combined influence certain factors may have. (Hint: plotting relationships can help in deducting early which factors do (not) influence satisfaction score). Make use of multivariate techniques.
19. In marketing there is a common phenomenon called ‘the long tail’ (look it up!). Are you able to observe similar patterns in this dataset with regard to the number of stops, number of daily travelers, etc.?
20. We can clearly observe a weekend-effect in the number of daily travelers? However, some stations experience more travelers during the weekend. Which stations are this? Plot them on a map of Belgium and try to deduct a reasoning to what causes these stations to observe an opposed effect?
21. Cluster stations based on week-weekend distribution of travelers (hint: consider rescaling). Do you observe similar clusters to the ones observed with regard to facilities-based clustering?

Overall, your presentation should make a comprehensive summary of the requested items. If you feel that one of the points requested above does not fit in your presentation, you can leave them out (or add it as extra slides). However, **you need to write code for all of the questions above**. There are a lot of extra aspects that can be looked at. You can include these in your presentation or coding if you think they may be of particular relevance or contain some unexpected results.

**NOTE**: We will also ask you to hand in your Python code, please structure it in a **readable** **format** using comments and numbering as below for each exercise. Any additional code you write, should be indicated as “extra”.

# Hints and guidelines

* New functions that have not been seen in class might be necessary, so use the Internet wisely.
  + Google
  + Stack Overflow
  + Quora
  + Etc.
* Handle missing values (delete them or find a way to integrate them in your analyses).
* **Do not limit yourselves to the questions asked above.** This will only result in very moderate grades. Try to think further and compile a value-adding business case.
* Search for relevant datasets on the internet if necessary. (Infrabel, NMBS, ...)
* Use plots and make your slides visually appealing.
* You can make use of the **Github Desktop application**. This application will make it very easy to upload your code to your UGent Github repo.
* Other information that you can find on the Internet can be useful to provide more detailed information.