INFOH423 – Project
Hack my Ride
Data Mining Project 2022/23

Justine Baret and Mahmoud SAKR
Justine.Baret@stib.brussels, mahmoud.sakr@ulb.be

Since you study in ULB, you must be a regular user of stib-mivb - the company providing public transport in Brussels. This already makes you a domain expert in this data challenge.

News: stib-mivb has an open data portal, on which the scheduling of the trips as well as the real-time vehicle arrival times are continuously published.

We have collected about 3 weeks of this data for you:

- The location of all vehicles every +-30 seconds, encoded in JSON. The JSON format is described at the end of this document
- Esri Shape files describing the map (lines and stops) of stib-mivb network, 23 September
- GTFS files containing the offline plan/schedule covering the same period of the vehicle location data, two snapshots 3 September and 23 September

One major concern for public transport is the satisfaction of passengers. In stib-mivb this is translated into punctuality and regularity of vehicle arrival times at stops. Punctuality is when the vehicle arrives at the stop in the time which is scheduled, as in the GTFS file. Note that the GTFS is encoding the schedule information which is printed on paper at the stops. For frequent lines, i.e., where the headway is lower than 12min, regularity is more relevant as a quantification of QoS than punctuality. For these lines, passengers tend to go to the stop without looking at the exact arrival times, as they know that they will have a maximum wait of a few minutes. For

these lines, the goal is to analyze whether the vehicles indeed arrive in the planned headway, rather than analyzing their exact arrival times. Notice that the planned headway per line changes depends on the hour of day. In peak hours, for instance, the headway will be smallest. In early and late night hours, the headway will be largest.

A quantitative measure that captures regularity is the EWT - Excess Waiting Time. A brief description of it can be found here, and you can yourself find many other resources explaining it. In your analysis, you can use this measure, and there is no restriction to use other measures if you find better.

Your team is challenged to analyze the given data, and provide quantitative insights about the quality of service. This overall all challenge is broken in the following tasks:

1. Distinguish the lines/periods/day/timegroups for which the QoS is assessed by punctuality, and those assessed by regularity.
2. Analyze the punctuality over the different lines, stops, and calendar days.
3. Analyze the regularity over the different lines, stops, and timegroups. Notice that, as a prerequisite to this analysis, you will probably need to identify the timegroups of different regularity per line, as explained during the lecture. If you find another way, you have all the freedom to apply your method.
4. In some cases, there are problematic segments that cause big delays or irregularity (bottlenecks). This delay / irregularity will then propagate to the rest of the line. Identify these lines/segments, and analyze if there is a pattern.
5. Think your own of a valuable analysis on this data

Deliverables

You should deliver a presentation of your work and results (as a .pdf) containing the following elements:

1. A cover page with the list of group members, including student ID

2. A description of dataset loading and preprocessing
3. A description of your data exploration activity; better accompanied with statistics, figures, screenshots, etc
4. A clear presentation of your solution (idea/algorithm) for every task. Special attention should be given to the correctness of the results
5. An interactive dashboard for visualizing the QoS KPIs that you extract in all the analysis tasks. The dashboard should be thought of to allow STIB data teams to visualize your analysis results of all the tasks 1-5, to visualize for certain lines, stops, neighborhoods, time periods, etc.

Evaluation

The evaluation jury consists of the two course instructors. STIB-MIVB representatives will also attend, depending on their availability. You will be asked to present your solution and defend it. The grading will consider the following factors:

- Your data management: loading, integration, exploration (5 points)
- For every analysis task (5 points each):
    - Your justification for the analysis pipeline and the parameter setting
    - Your presentation and interpretation of the results
- Your validation/evaluation of the obtained results in every task. Think of answers to questions like: how far are your results accurate? have you cross checked your results with some ground truth? What would limit the use of your solution in stib-mivb (5 points)

Vehicle location JSON file format

```
{"data":[{
  "time":"1632409236387",
  "responses":[
     {"lines": [
     {"lineId": "1",
     "vehiclePositions": [
          {"directionId": "8161",
          "distanceFromPoint": 1,
          "pointId": "8122"}, ...]}, "lineId": ...]
     },{"lines":...}
  ],}, {...}, ... ]}
```

This data has been collected by invoking the Vehicle Position Real-Time API of stib-mivb. Every 30 seconds, the API was called 9 times, each with 10 lines IDs. The `time` attribute is the time in milliseconds (unix epoch) at which the API was invoked. The `responses` array has the result of the 9 API calls. Every call returns for the given line IDs all their vehicle positions. Note that for one line, there are normally multiple vehicles at the same time, positioned across the route of the line.
A vehicle position, is a triple:

```
{"directionId": "8161",
 "distanceFromPoint": 1,
 "pointId": "8122"}
```

The `directionId` is the identifier of the terminal stop. The `pointId` is the identifier of the last stop traversed by the vehicle. The `distanceFromPoint` is the distance in meters between the vehicle and the last traversed stop.
The stop data (id, location, name, etc), as well as the routes of the lines are given in the Esri Shape files.