Python Project TIL6022
Research Proposal

Group 6

Date: 4-Oct-2024

Contributors:

- Thijs Daemen - 5289491
- Alene Hooiveld - 5310539
- Chris Juárez - 5171806
- Mathijs Markus - 5405416
- Niels van der Rijst - 5380162

Introduction

In this research the current capacity of the train network of the NS time schedule is studied. The objective is to explore notable features of the system and find the impact of the location of the train station,

For the analysis, data is used from an online database found through the following website: http://data.openov.nl/ . The data analysis in done for the scheduled train services for the week between 20 and 27 of September 2024.

Research question

The main research question of this study is: What are the capacities of the train network in the current NS time schedule?

This question will be answered by answering the following sub-questions:

- What is the difference in capacities of trains between the Randstad and outside the Randstad?
- How does the capacity differ between different days of the week?
- To what extend does the capacity when looking at train types, i.e. Sprinter vs. Intercity trains?

Data pipeline

The data that is used for the research is extracted from databases of the NS (https://data.ndovloket.nl/bezetting/ns/) and the online database http://data.openov.nl/. The first database is called OC_NS_20240920.csv where all the train operations between 20-09-2024 and 29-09-2024 are displayed. The data includes the operating day, line planning number, journey number, reinforcement number, timing link order, the code of the begin station and end station, occupancy, vehicle type and total number of coaches. To this data a column ‘Seats’ and ‘Occupied Seat’ is added. The data in ‘Seat’ is defined by number of coaches multiplied by the number of seats of a coach of a specific train type. Then the first three days are filtered out of the data in order to have a time span of a week.

The second database is called Trainservices.csv. The file consists of all routes that are in the current train schedule of the NS. It starts with a column with the start and end station, then a column with a code and a third column including all the train station where the train stops.

With these two datasets per train series a dataframe is made consisting of two consecutive train stations and the seat capacity in a week for that series. These dataframes are added to one dictionary that consists of the sum of the seat capacity of all series in a week for the consecutive train stations. An example can be seen in figure 1.

	From	To	Seats
0	Ut	Utvr	95148
1	Utvr	Utl	188892
2	Utl	Htn	188892
3	Htn	Htnc	377784
4	Htnc	Cl	377784
5	Cl	Gdm	377784
6	Gdm	Zbm	377784
7	Zbm	Ht	377784

Figure 1: Dictionary of the total seat capacity in a week

The seat capacity will be visualized by a map of the train network in The Netherlands. The train tracks will be given a color between red and blue. The more red the track the higher the seat capacity. The more blue the track is, the lower the seat capacity.

Contribution statement

Thijs Daemen

- Finding data, requesting API's, and starting with the code

Alene Hooiveld

- Defining the research questions and the objectives of the research, wrote the data pipeline.

Chris Juárez

- Defining Research questions, looking into streamlit for possible use in the project

Mathijs Markus

- Introduction, research question, set up GitHub repository for the project

Niels van der Rijst

- Used open data to make a list with all train services in the Netherlands