SNCF student project

Introduction and Data source

For this project, I will be working with data from SNCF, the National Company of the French Railways. This is internal, administrative data, provided by a state-owned company, and therefore fairly trustworthy. It is up to date: most of the data sets give information about the trains until June 2024.

SNCF website offers a variety of data sets on different topics, and I will combine several of them to be able to find relationships between different variables describing the passenger's transportation and characteristics of train stations.

I selected those data sets for my analysis:

- Monthly regularity for TGV (high speed trains – national and European), 2018-2024 (9599 rows) For each TGV line (from station A to station B), the number of trains planned for the given month, the number of trains that was on time/delayed/cancelled, average time, average delay…
- Number of passengers per year for each train station in France, 2015 - 2023 (3011 rows)
- Train stations and geographic position in GeoJSON format (2886 rows)

Data profile

Cleaning steps

TGV Delay data set

- Dropped 2 empty columns: "Commentaires annulations" and "Commentaire retard au départ"
- Renamed all columns - translation to English
- Column "Avg delay of all trains at departure": replaced 158 negative values with calculation based on other columns.
- Column "Avg delay of trains delayed at arrival": 2 negative values (2019-11 Montpellier – Paris Lyon and 2019-11 Nimes – Parys Lyon). Replaced with absolute value.
- Column "Avg delay of all trains at arrival": replaced 109 negative values with calculation based on other columns.
- Column "Avg delay of trains >15 (if trip has a flight concurrence)" had 2157 values over 15, which is 22% of incoherent value. Column not relevant to analysis: dropped.
- 73 rows where "Number of trips scheduled" AND "Avg trip length" = 0. All values from 2020, mostly April and May. Looks like disturbance from the pandemic. The problem is that for those rows, the "Number of trains cancelled" is not always 0, and it should not be higher than the number of trips scheduled. I changed to value of Number of trains cancelled" to 0.

Number of travelers data set

- Dropped the postal column as well as all the columns with the number of visitors of the stations who did not travel.
- Renamed all columns - translation to English.

Data understanding

With the data at hand, I'll be able to calculate the average delay and percentage of delayed trains/cancelled trains based on certain criteria (is the trip international or national, is the train Paris-bound or not, what region is it leaving from…)

In order to do this, I will derivate new variables (they will also help me reach the required number of categorical variables):

- "Paris bound?" With 3 different values: "Paris bound", "Leaving Paris", and "Province to province"
- "Departure region" and "Arrival region" can be deduced from the departure and arrival stations.
- "Busy station?" with 3 different categories, depending on the number of visitors/year given by the 2nd data set

Limitations and ethics

I'll only be considering TGV lines (high speed) and not the TER (regional ones), which is leaving a big part of the French rail traffic outside out the scope of the project. This could be considered a limitation.

Our main data set is very recent, so there will be almost no time lag to take into account. The second data set gives the number of travelers per year, it is only collected once a year, so that the latest data is from 2023, which is reasonable.

As for ethics, we're not dealing with any personal information, which leaves very little room for ethical issues to rise.

The data, collected automatically (as opposed to manually) by a government-owned organization, has virtually no potential for bias.

Questions to explore

What is the top 3 most punctual train stations in France? The 3 least punctual ones?

Is the average delay greater in some regions than in others?

Are trains leaving Paris more often delayed/cancelled than trains going to Paris? Than trains going from province to province?

Do train stations that receive a lot of visitors have tendency to have more delay on departure/arrival?

What are the train stations that received the most travelers in 2023 (including, and then excluding Paris) Are they the same as in 2015?

What is the proportion of trains delayed over 60 minutes in each region?

What was the impact of the 2020 pandemic on the French rail traffic?