COM 402 – Data Visualization: Milestone 2

Rail Runners - Giacomo Orsi, Francesco Salvi, Roberto Ceraolo

Live website here

1 Introduction

Our project aims to shed light on the efficiency of the Italian railway system, leveraging a novel dataset of historical train logs from Trenitalia over several months. While the dataset could lead to several exciting directions, such as studying connectivity across Italy in terms of travel times and rail availability, we decided to focus specifically on delays, because they are the main element associated with national trains in the popular perception. As a well-disciplined team of data scientists, common perception is however not enough, and we will rather approach that question with data: *are Italian trains really as bad as people think?*

2 Features

To answer our question in a clear and impactful way, we will structure our visualizations along two key axes: a **Stations** view, providing an immediate overview of the landscape across the whole country, and a **Train** view, focusing on individual trains to showcase advanced statistics and insights.

2.1 Stations view

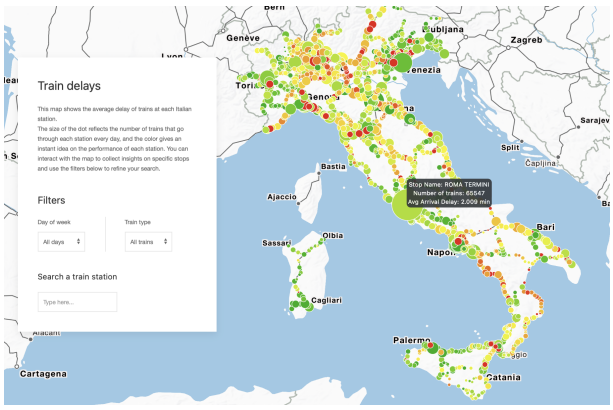

Figure 1: Station view

Core features The Stations view, shown in a sketch in Figure 1, will be the main landing point of our website, welcoming the users as they open it. They will face a full-screen map of Italy, with an overlay showing the main rail network and cities. On top of it, they will see a colored dot for each station, having a size proportional to the number of trains stopping there and a color that depends on the average delay across all the train stops. This view will give an intuitive overview of the national landscape, showing where the rail network is more or less connected and which areas have the worst punctuality on average. On the left side of the screen, a rectangular box will overlay the map. The box will provide basic instructions for the map and will include interactive components that will allow the users to filter by day of the week and type of train, with the map updating in real-time. Moreover, a search box will allow users to select a specific station and center the map on it, while showing detailed statistics about that stop.

Extra features If time allows, we would also want to show in the hover window at each station a small histogram with the average delay per day of the week, so that users can understand how much variation there is in weekdays vs weekends, which is a frequently debated point regarding the system's efficiency. Additionally, we would want to find a way to make smooth transitions when filtering or searching for a station, so that the update on the map is not sharp but rather gradual and appealing.

2.2 Train view

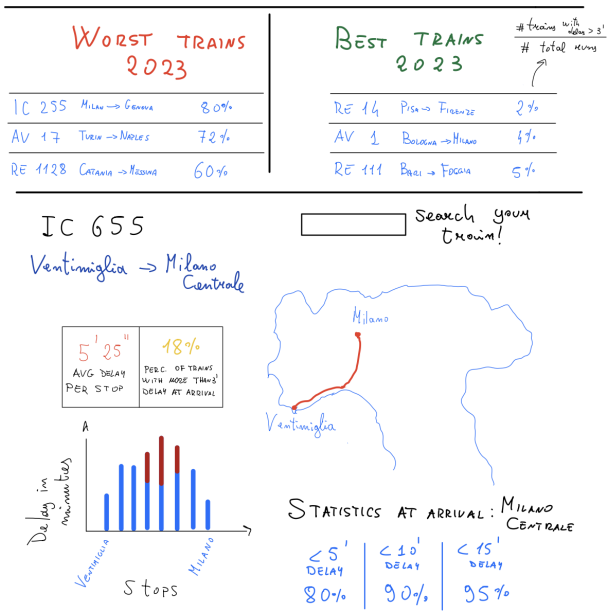

Figure 2: Train view

Core features The Train view, sketched in Figure 2, will further explore our dataset at the level of individual trains. The top part of the screen will be filled with a table of the worst and best trains across the period of interest, ranked by their percentage of delayed arrivals (where delayed = delay greater than 3 mins). The bottom part of the screen, instead, will feature an in-depth view of a single train, showing the path of the train on a smaller map and some aggregate statistics, including its average delay per stop and its percentage of delayed arrivals. Additionally, we will show a barplot of the average delay per stop, where clicking on a bar shows statistics such as the percentage of arrivals delayed more than 3/5/10 minutes. Users will be able to change the train currently displayed by either selecting one of the trains in the top table or by manually searching for a train number in the dedicated search bar.

Extra features If time allows, we aim to connect the statistics of train stops both with the stops histogram and with the map, so that users will be able to click on a station on either of those two elements and see its relative statistics. Also, thanks to our tools, we would like to analyze the data to draw insights and prepare a data story. The following is a non-exhaustive list of what we would like to investigate about: differences in the quality of service in the north vs the south of Italy, between the types of trains (regionals, inter-regionals, high speed trains), between weekdays and weekends and between peak and non-peak times.

3 Tools

- Website: the website will be hosted on GitHub pages. It will be a combination of HTML/CSS/JS files. Jekyll will be used to manage all those files.
- UI libraries: UIKit, a lightweight CSS/JS library is used to model the UI of the website
- Data Visualizations: D3.js will manage the visualization of the dots of each station on the maps and to generate the histograms mentioned above.
- Maps: MapBox GL is used to display maps on the website. MapBox allows extensive customizations on the style of the map, which we will need in order to display a map that highlights railways.
- Python: it was crucial for the initial data cleaning, wrangling, and exploration. We used PySpark to handle the large amount of data and Pandas to clean and have a first grasp of the possible analysis to be done.

For the main visualizations, we will make intensive use of Lectures 2-5, covering the basics of Javascript and d3.js. Given the focus of the project on maps, will use Lecture 8 for the main geographic plots. Finally, we will use Lectures 7 and 12 to guide our storytelling and to make sure to avoid bad practices across all our visualizations, and Lecture 6 to decide how to best encode insights for each data point.