Data collection and Storage Project

Subject: Railway

MATHIEU Antoine

ODIC Nathan

GONG Kaiyuan

SIDAHMED Sidi

Table des matières

I- Subject and problematic 3
II- Structure of ours DOE 3
III- Web scraping of the data. 4
IV- Storage 6
V- Conclusion 7

I- Subject and problematic

We decided to choose the subject of railway because for us it was interesting, and it has a lot of possibilities. We started by defining our question, we decided to focus on the forecast of the SNCF and how relevant it was. Then we ended up with this question:

What is the reliability of the SNCF forecast for a delayed train and what effect can be easily detected by the SNCF to forecast a delay?

The goal of this question is to see how the forecast of the SNCF is relevant and what type of delay can be easily detected in advance and what type of delays are unexpected.

II- Structure of ours DOE

For our DOE we started by defining our question then we for the population of our experiment we decided to focus French train and especially SNCF's train (TGV, TER and Intercités). We chose to delete the RER and Transiliens trains because the SNCF API names the stations crossed by the RER/Transiliens and the other trains differently (e.g., "Paris Gare de Lyon - Hall 1&2" and "Paris Gare de Lyon"), which makes it impossible to clearly recognize the stations in the dataset.

For the sample size we tried to get as many as possible sample, at the end we have approximately for our 3 databases:

- 4000 trains at departure
- 1000 trains at arrival
- 1000 delayed trains

We should have collected much more data but the 2 days of general strike in France caused a lot of train cancellations and therefore very few delays. However, we are only interested in late trains, and we remove the cancelled trains from our data, hence the small amount of data. It should be noted that the data retrieved through the API is live and therefore disappears after the end of the train journey, making it impossible to retrieve previous data.

Our DOE will have two types of measure. It will measure the forecast delay of a train and its real delay. Then we defined different factors to see how they influence the forecast. We choose the departure station and arrivals station of the train, all its stops in different station. The hour of departure and arrivals. The idea behind these factors is to see is the time of departure, stops and the length of the journey have an impact on the forecast. All these factors have different levels detailed below:

- Departure Station levels: "station name" (we use the 3220 train station referend by the SNCF)
- Arrivals Station levels: "station name"

- Stops levels: [“stop name 1”,…,”last stop name”]
- Departure levels hour: hh:mm:ss
- Arrival levels hour: hh:mm:ss
- Departure day: dd-mm-yyyy
- Cause of delay: “type of delay”

Our hypothesis that we want to prove is:

H0:” the predicted delay or not of a train is true”

H1:” the predicted delay or not of a train is false”

Finally, we decided to choose a Latin square because we a lot of factors and they a not interacting we’ve each other.

III- Web scraping of the data.

Now that our DOE is defined, we searched for API to web Scrap the data. SCNF offers an API with a version 1 and 2. However, individuals like us only have access to the first version, but which is already very complete.

For departing and arriving train datasets:

We make a first request to this API to get the list of the stations in France and the code associated with them. This allows us to make a loop to inspect the trains at the departure and arrival in all the French stations. Considering the length of the query to search all the French stations (3h for the 3220 stations), we have chosen to focus on the 25 largest French stations.

Once this query is done, we filter the trains by deleting the RER and Transiliens for the reasons expressed in the previous section. Then, we delete the trains that have been cancelled, replaced or diverted to keep only the late trains.

After some processing steps, we know the destination of the train, its type, number, trip ID, hours and delay, but we need some additional information like the origin of the train, the stops and the cause of the possible delay. To do this, we make a new request to the API for each of the trains obtained through their trip ID.

This step is only for one station, so it remains to make a loop so that the search is done in each of the 25 selected stations.

For the delay dataset:

For this dataset, we make a request to the API to retrieve all trains that are late since a certain date. Here too we keep only the trains that have been delayed.

For each of the late trains, we keep:

- The day when the train left.
- The station of departure and arrival
- Expected and actual arrival time
- The time of departure
- Delay
- The cause of the delay

	Provenance	Destination	Train	Numéro	Jour	Départ (réel)	Départ (prévu)	Retard (min)	Cause	Arrêts
0	Bordeaux Saint-Jean	Agen	TER NA	866831	09-03-2023	15:33:00	15:33:00	0		[Beautiran, Cérons, Langon, La Réole, Marmande...
1	Hendaye	Paris - Montparnasse - Hall 1 & 2	TGV INOUI	8540	09-03-2023	15:46:00	15:46:00	0		[Paris - Montparnasse - Hall 1 & 2]
2	Bordeaux Saint-Jean	Arcachon	TER NA	866341	09-03-2023	16:04:00	16:04:00	0		[Pessac, Pessac Alouette, Gazinet Cestas, Marc...
6	Bordeaux Saint-Jean	Paris - Montparnasse - Hall 1 & 2	TGV INOUI	8488	09-03-2023	16:49:00	16:19:00	30	Régulation du trafic	[Libourne, Angoulême, Poitiers, Châtellerault,...
10	Bordeaux Saint-Jean	Paris - Montparnasse - Hall 1 & 2	TGV INOUI	8418	09-03-2023	17:08:00	17:08:00	0		[Massy TGV, Paris - Montparnasse - Hall 1 & 2]
16	Bordeaux Saint-Jean	Paris - Montparnasse - Hall 1 & 2	TGV INOUI	8420	09-03-2023	17:46:00	17:46:00	0		[Paris - Montparnasse - Hall 1 & 2]
20	Bordeaux Saint-Jean	Arcachon	TER NA	866351	09-03-2023	18:35:00	18:35:00	0		[Pessac, Pessac Alouette, Gazinet Cestas, Marc...
26	Libourne	Arcachon	TER NA	866203	10-03-2023	06:35:00	06:35:00	0		[Pessac, Pessac Alouette, Gazinet Cestas, Marc...
27	Bordeaux Saint-Jean	Sarlat	TER NA	865704	10-03-2023	07:41:00	07:41:00	0		[Cenon, Libourne, Saint-Émilion, Castillon, Vé...
28	Bordeaux Saint-Jean	Paris - Montparnasse - Hall 1 & 2	TGV INOUI	8404	10-03-2023	07:45:00	07:45:00	0		[Massy TGV, Paris - Montparnasse - Hall 1 & 2]

Picture of the departure Dataframe for the Bordeaux train station the 09/03/2023 at 15h13

	Provenance	Destination	Train	Numéro	Jour	Arrivée (réelle)	Arrivée (prévue)	Retard (min)	Cause	Arrêts
0	Langon	Bordeaux Saint-Jean	TER NA	866744	09-03-2023	15:17:00	15:17:00	0		[Langon, Preignac, Cérons, Podensac, Portets, ...
1	Hendaye	Paris - Montparnasse - Hall 1 & 2	TGV INOUI	8540	09-03-2023	15:40:00	15:40:00	0		[Hendaye, Saint-Jean-de-Luz - Ciboure, Biarrit...
2	Nantes	Bordeaux Saint-Jean	Intercités	3833	09-03-2023	16:07:00	16:07:00	0		[Nantes, La Roche-sur-Yon, Luçon, La Rochelle,...
3	Paris - Montparnasse - Hall 1 & 2	Bordeaux Saint-Jean	OUIGO	7653	09-03-2023	16:09:00	16:09:00	0		[Paris - Montparnasse - Hall 1 & 2, Poitiers, ...
4	Paris - Montparnasse - Hall 1 & 2	Tarbes	TGV INOUI	8574	09-03-2023	16:14:00	16:14:00	0		[Paris - Montparnasse - Hall 1 & 2, Massy TGV,...
5	Paris - Montparnasse - Hall 1 & 2	Bordeaux Saint-Jean	TGV INOUI	12265	09-03-2023	16:14:00	16:14:00	0		[Paris - Montparnasse - Hall 1 & 2, Massy TGV,...
6	Paris - Montparnasse - Hall 1 & 2	Toulouse Matabiau	TGV INOUI	8509	09-03-2023	17:15:00	17:15:00	0		[Paris - Montparnasse - Hall 1 & 2, Bordeaux S...
7	Arcachon	Libourne	TER NA	866246	09-03-2023	17:25:00	17:25:00	0		[Arcachon, La Teste, La Hume, Gujan-Mestras, L...
8	Paris - Montparnasse - Hall 1 & 2	Bordeaux Saint-Jean	TGV INOUI	8485	09-03-2023	17:37:00	17:37:00	0		[Paris - Montparnasse - Hall 1 & 2, Saint-Pier...
9	Marseille - Saint-Charles	Bordeaux Saint-Jean	Intercités	4760	09-03-2023	17:38:00	17:33:00	5	Prise en charge de clients en correspondance	[Marseille - Saint-Charles, Arles, Nîmes Centr...

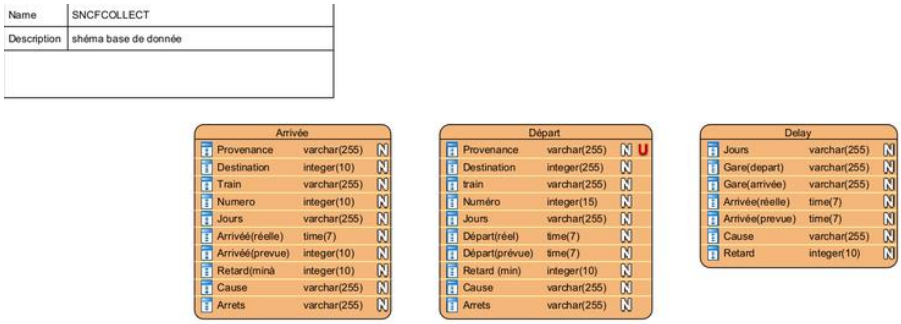
Picture of the arrival Dataframe for the Bordeaux train station the 09/03/2023 at 15h13

	Jour	Gare (départ)	Gare (arrivée)	Arrivée (réelle)	Arrivée (prévue)	Départ (prévu)	Retard (min)	Cause
14	08-03-2023	Montpellier Saint-Roch	Valence TGV Rhône-Alpes Sud	07:54:00	07:04:00	05:50:00	50	Mise à quai tardive en gare origine
17	08-03-2023	Maubeuge	Charleroi-Central	20:53:00	20:48:00	19:53:00	5	Incident sur un réseau ferré étranger
19	08-03-2023	Strasbourg	Bâle - SBB	21:13:00	21:08:00	19:51:00	5	Régulation du trafic
104	08-03-2023	Givet	Charleville-Mézières	19:00:00	18:25:00	17:20:00	35	Réutilisation d'un train
105	08-03-2023	Orléans	Tours	18:27:00	18:22:00	17:00:00	5	Présence d'individus sur les voies
...	...	...	...	...	...	...	...	...
34170	08-03-2023	Dinan	Lamballe	19:22:00	19:17:00	18:10:00	5	Embouteillage
34226	08-03-2023	Amiens	Laon	20:05:00	20:00:00	18:25:00	5	Régulation du trafic
34269	08-03-2023	Verdun	Metz Ville	13:02:00	12:57:00	10:37:00	5	Embouteillage
34280	09-03-2023	Lille Flandres	Tournai	11:21:00	10:36:00	10:08:00	45	Réutilisation d'un train
34311	09-03-2023	Toulon	Marseille - Saint-Charles	16:30:00	15:55:00	14:50:00	35	Régulation du trafic

Picture of the delay Dataframe

IV- Storage

Once the date is collected, we decided to store it in a NoSQL database using MongoDB. We create 3 table arrival, departure, and delay. Below we have a diagram showing the structure of our NoSQL Database.



The idea behind it is that with the day, hour, train departure and arrival station, it's easy to correlate the arrival and departure table to the delay table. So, the journey of the train with all of his stop and their schedule can be easily determine.

V- Conclusion

The project allowed us to develop our web scraping skills. Indeed, we had to understand the API of the SNCF and get the data we needed. Our main difficulty was the pre-processing of the data, which data are useful and how to transform them to make them usable. This difficulty allowed us to see how to use pandas to format the data. Once this difficulty was overcome, we realized that the query was much too long (3hour to collect only 5 trains per station) because it was going through all 3220 stations. So, we reduced the number of stations to compose our dataset by keeping only about 50 stations.

To conclude we create a dataset with 3 table that can be easily connected to each other. The structure of the data will allow us to answer to our question.