# Rapport Projet DataWarehouse2 : Analyse des Retards de Trains et Météo en Décembre 2024

Auteur: HABBAK Taous  
Date: 19/01/2025

## Introduction

Dans ce projet, j'ai analysé les retards des trains à la gare de Lyon en décembre 2024 en croisant les données des horaires de départ des trains avec les conditions météorologiques de ce mois. L'objectif principal était d'évaluer l'impact des conditions climatiques sur la ponctualité des trains et de visualiser ces résultats à travers des graphiques interactifs. Les données utilisées ont été générées par une **IA**, ça m'a permis de créer les ensembles de données nécessaires à l'analyse.

## Objectifs du Projet

* **Analyser l'impact des conditions météorologiques** (température, précipitations, humidité, vent) sur les retards des trains.
* **Visualiser les données** à l'aide d'une interface interactive pour une meilleure compréhension des relations entre les retards et la météo.
* Identifier les principales **causes des retards**, en particulier celles liées à la météo, et évaluer leur poids par rapport à d’autres causes.

## Méthodologie

Le projet suit les étapes classiques d'un processus **ETL** (Extraction, Transformation, Chargement), et inclut également une phase de visualisation des données à l'aide de **Dash** et **Plotly**. Voici les différentes étapes du processus.

### 1. Extraction des Données

Pour commencer, j'ai extrait les données de deux fichiers CSV distincts :

* **Horaires des trains** : Ce fichier contient les informations relatives aux départs des trains (heure de départ prévue et effective).
* **Météo** : Ce fichier fournit des informations météorologiques quotidiennes pour le mois de décembre 2024, incluant la température, les précipitations, l'humidité et la vitesse du vent.

Ces données ont été générées par une **IA Gen**, un générateur de données synthétiques, ce qui permet d’obtenir des ensembles de données réalistes tout en évitant l’utilisation de données réelles ou sensibles.

J'ai nettoyé les fichiers en supprimant les doublons et les valeurs manquantes dans la colonne date, pour m'assurer que les données soient complètes et fiables.

### 2. Transformation des Données

Une fois les données extraites, je les ai transformées afin de pouvoir effectuer une analyse approfondie :

* **Remplacement des valeurs manquantes** : L’heure de départ effective des trains a été utilisée pour remplacer les valeurs manquantes dans cette colonne.
* **Fusion des données de train et météo** : J'ai fusionné les données des horaires des trains et de la météo sur la colonne date, ce qui m’a permis d’associer chaque train aux conditions météorologiques correspondantes.
* **Calcul des retards** : J'ai calculé la durée des retards en soustrayant l'heure de départ prévue de l'heure de départ effective. Les retards de 10 minutes ou plus ont été considérés comme significatifs.
* **Identification des mauvaises conditions météorologiques** : J'ai jugé que les conditions météorologiques étaient mauvaises si :
  + La température était inférieure à 0°C.
  + Les précipitations étaient supérieures à 10 mm.
  + L'humidité dépassait 80%.
  + Le vent soufflait à plus de 10 km/h.
* **Ajout des causes des retards** : Les causes des retards ont été déterminées comme suit :
  + Si un retard significatif (>= 10 minutes) coïncidait avec des conditions météorologiques défavorables, la cause était marquée "Mauvais Temps".
  + Sinon, la cause était marquée "Autre Cause".
* **Nettoyage des données** : Les données ont été filtrées pour ne conserver que celles ayant des retards significatifs et les colonnes pertinentes.

### 3. Chargement des Données

Ensuite, j'ai chargé les données brutes (horaires des trains et météo) ainsi que les données transformées dans une base de données **SQLite**. Cela m'a permis de stocker les informations de manière structurée et de faciliter leur consultation pour des analyses futures.

### 4. Visualisation des Données

L’analyse des données a été complétée par la création de graphiques interactifs à l’aide de **Dash** et **Plotly**. J'ai créé trois graphiques principaux :

* **Température quotidienne** : Un graphique linéaire montrant l’évolution de la température au cours du mois de décembre 2024.
* **Durée moyenne des retards par cause** : Un graphique à barres illustrant la durée moyenne des retards en fonction des causes (mauvais temps ou autres causes).
* **Nombre de trains par jour** : Un graphique à barres montrant le nombre de trains qui ont circulé chaque jour en décembre 2024.

### 5. Résultats

L’analyse a révélé que les **conditions météorologiques défavorables** sont la principale cause des retards des trains en décembre 2024. En effet, les jours de **mauvais temps** (températures sous zéro, précipitations élevées, humidité et vent forts) ont montré une **augmentation significative des retards**, en raison de perturbations dues au gel des voies, à des risques accrus de défaillances techniques et à des mesures de sécurité supplémentaires.

Les **autres causes de retard**, comme des problèmes techniques ou des événements imprévus, ont été moins fréquentes et ont eu un impact moins significatif sur la durée des retards.

### 6. Conclusion

Cette analyse montre clairement que la **mauvaise météo** a un impact majeur sur la ponctualité des trains à la gare de Lyon en décembre 2024. Les conditions climatiques extrêmes ont causé les retards les plus importants, et ces résultats mettent en évidence l'importance de tenir compte des prévisions météorologiques dans la gestion du trafic ferroviaire. Les retards dus à des facteurs autres que la météo étaient moins fréquents et moins significatifs, ce qui souligne encore l'importance des conditions climatiques dans la gestion des retards.