5ISS - Traitement et analyse de données

La SNCF est-elle toujours en retard ?

*"*Tum tum tudum* Chers voyageurs, nous vous informons que suite à un problème sur voie, le train à destination de Toulouse Matabiau sera retardé d'une heure. Merci de votre compréhension."* Qui n'a jamais entendu une telle annonce en gare ? C'est bien connu, les trains sont toujours en retard… Toujours, vraiment ? C'est le mystère que nous tenterons d'élucider dans ce rapport. Notre analyse vise à étudier les retards sur les lignes de TGV opérées par la SNCF en France, entre 2015 et 2020. Le dataset utilisé est disponible sur le site de la SNCF et recense la régularité mensuelle des TGV par liaisons. L'étude repose sur un grand nombre de critères, et prend en compte le nombre de trains prévus, le nombre de trains annulés, le moment de l'année le plus propice à un retard ou une annulation, la durée moyenne du retard par ligne et les causes majoritaires de retard. Nous tenterons aussi d'établir les lignes 'pire cas' et 'meilleur cas' en termes de retard en France, afin de vous conseiller pour vos trajets.

1. Annulations des TGV par année

Dans cette première partie, nous allons nous intéresser aux annulations par année pour comprendre si les TGV sont majoritairement annulés à une période spécifique de l'année.

Tout d'abord, analysons le diagramme de la Figure 1, intitulé "Nombre de trains prévus par an". Comme son nom l'indique, ce diagramme représente en ordonnée le nombre de TGV prévus par la SNCF sur la période allant de 2015 à 2020. On peut noter que le nombre est relativement constant d'année en année, à l'exception de l'année 2020 car les observations s'arrêtent en juin 2020.

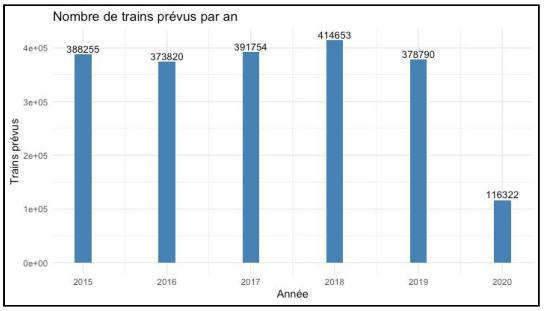

Figure 1 : Nombre de trains prévus par an

Maintenant, intéressons-nous au nombre de trains annulés à partir du diagramme de la Figure 2 intitulé "Nombre de trains annulés par an".

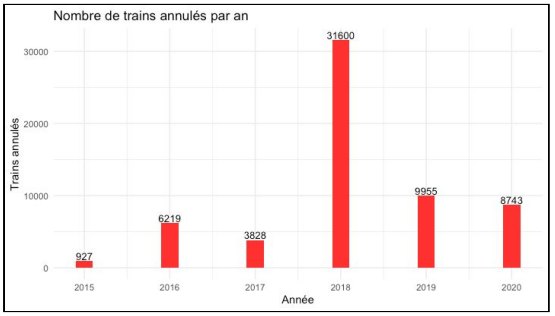

Figure 2 : Nombre de trains annulés par an

Nous pouvons noter que le nombre de trains annulés est particulièrement élevé durant l'année 2018. Après recherche, nous avons compris que cela était lié à l'important mouvement de grève des cheminots de la SNCF.

Enfin, le plus important dans notre démarche est d'analyser le pourcentage d'annulation de TGV, c'est-à-dire le rapport entre le nombre de trains annulés et le nombre de trains prévus.

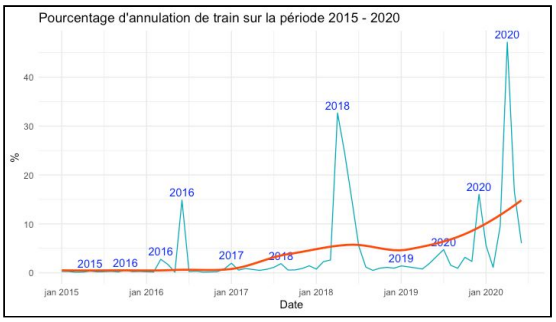

Figure 3 : Pourcentage d'annulation sur la période 2015 - 2020

D'une part, cette dernière analyse présentée dans la Figure 3 permet de confirmer un fort taux d'annulation en 2016 et en 2018 dû à des mouvements de grèves. D'autre part, cette analyse permet de révéler un très fort taux d'annulation durant l'année 2020 lié à la crise sanitaire, taux qui n'était pas autant visible dans les diagrammes précédents. Il est donc toujours important d'analyser des données en les mettant en perspective avec des données de référence.

Désormais, essayons de déterminer la période de l'année où le taux d'annulation est le plus important. Pour cela, nous traçons le pourcentage d'annulation de chaque année dans le diagramme de la Figure 4.

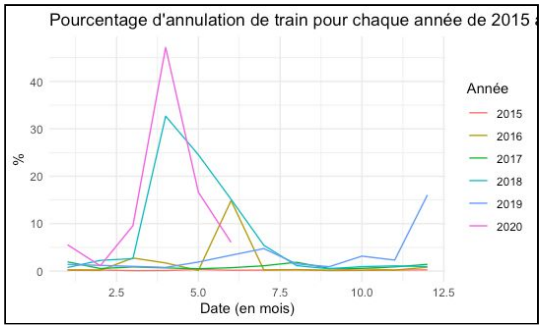

Figure 4 : Pourcentage d'annulation pour chaque année de 2015 à 2020

Nous pouvons voir que la période de mars à juin est une période de l'année particulièrement sujet à des annulations, tandis que la période d' août à octobre représente des taux d'annulation très faibles.

Ce dernier diagramme présenté dans la Figure 5 représente la moyenne des taux d'annulation sur toutes les années et permet de confirmer nos précédentes observations.

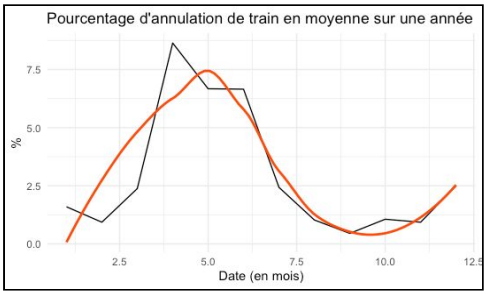

Figure 5 : Pourcentage d'annulation en moyenne sur une année

2. Analyse des principales causes de retard et de leur évolution

Afin de comprendre "pourquoi les trains sont toujours en retard", nous avons décidé d'analyser les causes de retard de tous les TGV opérés par la SNCF entre Janvier 2015 et Juillet 2020. Sur la Figure 7, chaque graphique représente le pourcentage de trains en retard pour une certaine cause, par année, avec en rouge la moyenne sur toutes les années. La Figure 6 présente les mêmes résultats sous forme de diagramme en étoile.

Ainsi on peut voir qu'en moyenne, les retards sont principalement liés à des travaux ou de la maintenance, ou à des conditions externes telles que la météo ou les mouvements sociaux. A elles seules, ces deux causes représentent plus de 50% des retards.

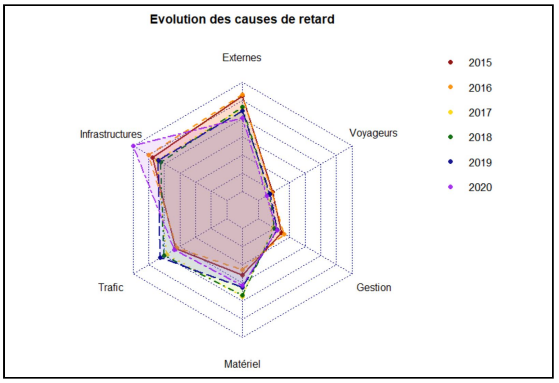

Figure 6 : Répartition des causes de retard sous forme de diagramme en étoile

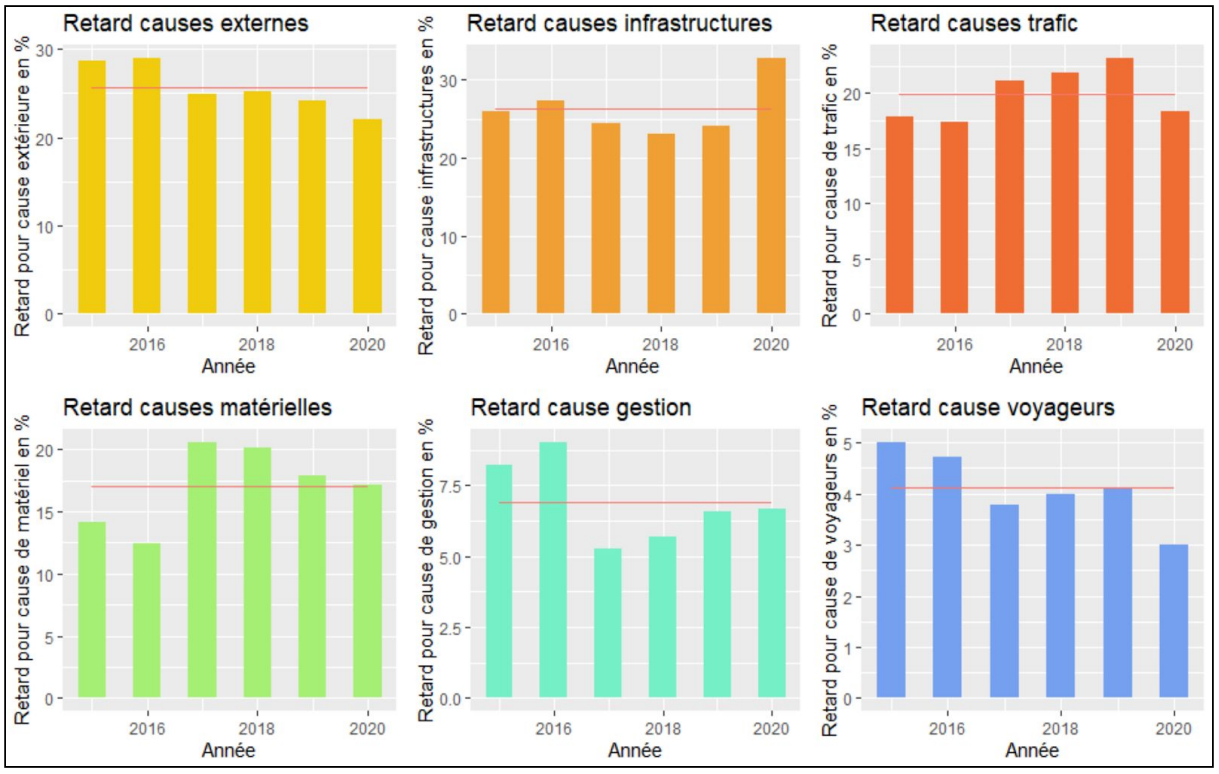

Figure 7 : Répartition des causes de retard en pourcentage (relativement au total des trains en retard) par année

On va maintenant s'intéresser à la variation des causes pour certaines années, qui peuvent souvent être reliées à l'actualité.

En 2020, on voit une augmentation de retard pour causes infrastructures, et on trouve sur le site de la SNCF: "En 2020, plus de 1 650 chantiers sont programmés et 6,2 Mrds€ seront investis dans le réseau. Ainsi 1 050 km de voies seront renouvelés et 500 aiguillages remplacés." En 2020, on remarque aussi que les retards pour causes voyageurs sont 1% en dessous de la moyenne pourtant assez stable, ce qui peut être mis en parallèle avec la baisse de fréquentation due à la pandémie de Covid-19.

On peut s'intéresser au retard par gare, en prenant par exemple une grande gare qui puisse être concernée par des évènements type pannes, gestion ou grèves.

En analysant les pics de retard en Figure 8, on peut se rendre compte encore une fois de l'impact des mouvements sociaux. Par exemple, des mouvements de grève ont été conduits en Juillet/Août 2018 ou encore en Mai 2016.

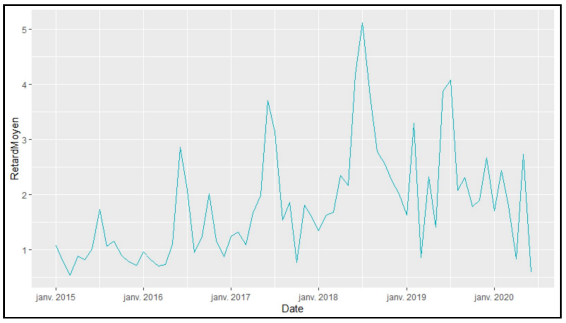

Figure 8 : Retard moyen en fonction de la date entre 2015 et 2020 pour la gare de Paris Lyon

3. Les pires cas et meilleurs cas: lignes à éviter et lignes à privilégier

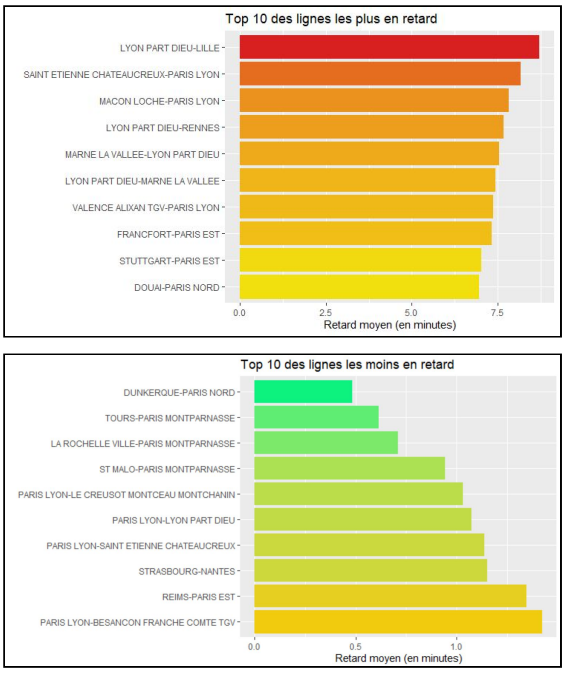

Figure 9 : Diagramme en barre présentant les 10 lignes les plus en retard et les plus à l'heure à l'arrivée

On s'intéresse maintenant aux lignes de TGV les plus en retard et les plus à l'heure opérées par la SNCF. Grâce à une analyse du dataset, nous avons pu établir un top 10 présenté en Figure 9. On remarque que les mauvais élèves sont Lyon Part Dieu et Paris (toutes gares), qui apparaissent respectivement 4 et 6 fois au classement des pires lignes. Ce résultat semble cohérent avec l'analyse précédente qui montrait que 50% des retards sont liés à des causes d'infrastructures ou de mouvements sociaux/colis suspects. Les grandes villes sont donc les plus concernées.

Mais Paris (toutes gares) apparaît aussi dans la liste des bon élèves, et ce 9 fois sur 10 ! On peut remarquer que ce sont des liaisons très courtes entre Paris et ses environs (au sens large) qui sont les plus à l'heure.

4. Analyse du retard en fonction de la durée du trajet

Suite à l'analyse de la Figure 9, on remarque que les lignes dans le rouge niveau retard ont tendance à être les plus longues, et les lignes en vert les plus courtes niveau distance parcourue (directement liée au temps de trajet). On se demande donc s'il existe une corrélation entre le retard d'un train et la distance qu'il parcourt. Grâce à une interpolation linéaire sur l'ensemble des données de retard des TGV, on se rend compte qu'en moyenne le retard à l'arrivée dépend bien de la durée moyenne du trajet. Cela paraît assez logique, car plus la distance parcourue est grande plus les incidents pouvant être rencontrés en route sont nombreux, impactant ainsi la durée du trajet.

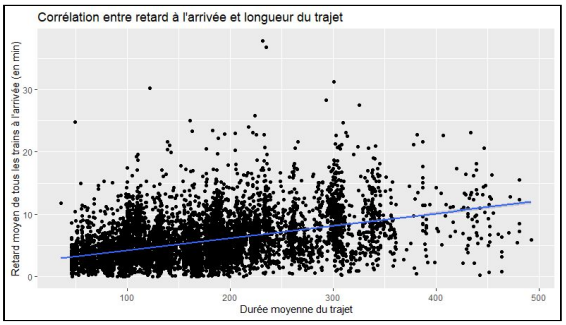

Figure 10 : Nuage de point présentant la corrélation entre le retard moyen à l'arrivée et la durée moyenne du trajet. En bleu, droite de régression linéaire avec marges d'erreur[1] *pour ces données.*

Finalement, on s'intéresse au temps d'attente d'un passager sur les quais, c'est-à-dire le retard moyen qu'un train aura au départ. Les courbes ne sont pas répétées, car elles montrent des résultats en tous points similaires à ceux de la Figure 9. Néanmoins, on peut légitimement se demander si la durée du trajet impacte aussi le retard qu'un train aurait au départ (bien qu'il n'y ait aucune cause évidente à cela). Grâce à une interpolation linéaire sur l'ensemble des données, on se rend

[1] On note que la marge d'erreur est si faible qu'elle n'est pas visible sur un graphique aussi peu zoomé.

compte sur la Figure 11 que notre intuition était la bonne: on ne peut pas établir de corrélation directe entre le départ au départ et la durée moyenne du trajet, la droite est constante à une valeur en y de 4 minutes, ce qui correspond au retard moyen de tous les TGV toutes durées et trajets confondus.

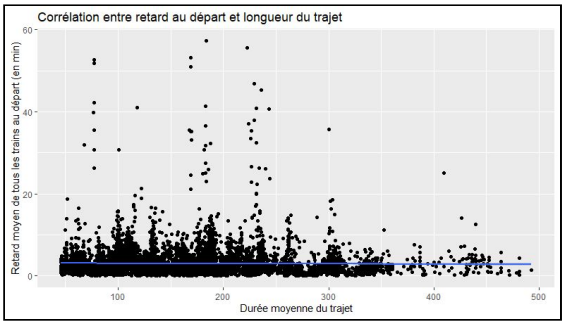

Figure 11 : Nuage de point présentant la corrélation entre le retard moyen à l'arrivée et la durée moyenne du trajet. En bleu, droite de régression linéaire avec marges d'erreur pour ces données.

5. Conclusion

**Chers passagers, nous espérons que vous avez apprécié votre trajet à bord de nos TGV, la SNCF vous remercie*.*

Finalement, les TGV sont relativement ponctuels avec une moyenne de 4 minutes de retard en moyenne tous trajets confondus. Nous avons identifié des facteurs aggravants: plus un trajet est long, plus les trains sont sujets à un retard à l'arrivée. Néanmoins ce facteur n'influe pas sur la ponctualité au départ: même pour un long trajet, vous ne devriez pas attendre plus longtemps sur les quais.

Nous avons aussi pu démontrer l'évidence: les gares les plus fréquentées sont aussi les plus sujettes aux annulations et aux retards. En cause: la gestion des infrastructures et du matériel roulant qui peut devenir un vrai casse-tête à une telle échelle, et les évènements type mouvements sociaux ou colis suspects.

Enfin, si vous voulez prendre un TGV sans avoir trop peur d'une annulation de dernière minute, nous vous conseillons d'organiser vos voyages durant la période du mois d'août au mois d'octobre et surtout pas durant la période du mois de mars au mois de juin !