

Temps de réponse de la Brigade des pompiers de Londres



LONDON FIRE BRIGADE

AVR24_continu_DA

Paul DEVIN
Thomas GARRIGUE
Marie HERBERT
Namita KALA

SOMMAIRE

ÉTAPE 1 - Exploration, data visualisation et pre-processing des données.....	5
I) Contexte du projet.....	6
1. D'un point de vue technique.....	6
2. D'un point de vue économique.....	7
II) Objectifs du projet.....	8
1. Principaux objectifs du projet.....	8
2. Problématique.....	9
3. Composition et maturité de l'équipe projet.....	9
4. Planning du projet - voir document Calendrier de projet-AVR24_CDA_Pompiers.pdf.	10
III) Cadre du projet.....	11
1. Jeux de données utilisés.....	11
a. Détail des incidents depuis 2009.....	11
b. Détail des interventions depuis 2009.....	12
2. Accessibilité et volumétrie.....	13
a. Détail des incidents depuis 2009.....	13
b. Détail des interventions depuis 2009.....	14
IV) Zoom sur les données.....	15
1. Pertinence des variables.....	15
2. Variable cible.....	16
3. Particularités remarquables du jeu de donnée.....	16
4. Données limitantes.....	18
V) Exploration et analyse des données avec DataViz'.....	19
1. Relation entre les variables explicatives et cibles.....	19
2. Distributions des données.....	20
a. Brainstorming de la Dataviz.....	20
b. Data visualisation.....	20
i. Analyse du temps d'intervention.....	20
ii. Analyse de la séquence de temps.....	22
iii. Analyse du lieu d'interventions.....	24
iv. Analyse du type d'interventions.....	27
v. Analyse du volume d'interventions par année, mois, jour, heure.....	31
vi. Analyse du nombre d'équipes et de pompes.....	33
vii. Analyse des temps de retards.....	35
3. Analyses statistiques.....	36
VI) Nettoyage et Pre-processing.....	38
1. Apprentissage supervisé ou non supervisé?.....	38
2. Régression ou classification?.....	38
3. Processus de Nettoyage et de traitement des données.....	38
a. Identification des doublons.....	38
b. Identification des valeurs manquantes.....	38
c. Variables explicatives conservées.....	39
i. Jeu de données "Incidents"	39

ii. Jeu de données "Mobilisations".....	40
d. Séparation des types de variables.....	40
Variables numériques :.....	40
Variables catégorielles :.....	40
4. Transformation des données.....	41
a. Jeu de données "Incidents".....	41
b. Jeu de données "Mobilisations".....	41
VII) Conclusion et projection sur la partie modélisation.....	42
 ETAPE 2 - Modélisation.....	43
I) Nos étapes d'itérations de modélisation.....	44
II) Interprétation et optimisation des modèles de Machine Learning.....	45
1) Première approche : interprétation.....	46
a) Avec les variables "TurnoutTimeSeconds", "TravelTimeSeconds", "NumCalls", "PumpMinutesRounded".....	46
b) SANS les variables hyper-correlées à la variable cible : "TurnoutTimeSeconds", "TravelTimeSeconds".....	47
c) SANS les variables hyper-correlées à la variable cible : "FirstPumpArriving_AT" et "SecondPumpArriving_AT".....	47
2) Seconde approche : optimisation.....	47
a) Feature engineering et pré-processing.....	47
b) Gestion des outliers corrélées à la variable cible.....	49
c) Approche avec l'ACP.....	50
d) Encodage des variables cycliques.....	50
e) Sélection aléatoire des lignes.....	51
3) Comparaisons des métriques des modèles testés.....	51
4) Analyse graphique des modèles testés.....	51
IV) Conclusions scientifiques et métiers de la modélisation.....	54
Conclusion scientifique.....	54
Conclusion métier.....	54
 ETAPE 3 - Conclusions projet.....	56
I) Difficultés rencontrées lors du projet.....	57
II) Bilan.....	58
1. Contribution principale dans l'atteinte des objectifs projet.....	58
2. Résultats obtenus comparés au benchmark initial.....	58
3. Atteinte des objectifs.....	59
III) Suite du projet.....	60
LISTE DES FIGURES.....	61
BIBLIOGRAPHIE.....	62
Données.....	62
Documents.....	62
Sites Internet.....	62

ÉTAPE 1 - Exploration, data visualisation et pre-processing des données.

I) Contexte du projet

1. D'un point de vue technique

Ce projet d'analyse des temps de réponse de la brigade des pompiers de Londres s'inscrit dans le but d'utiliser l'ensemble des librairies Python apprises lors de notre formation de Data Analyst.

La brigade des pompiers de Londres est le service d'incendie et de sauvetage le plus actif du Royaume-Uni et l'une des plus grandes organisations de lutte contre l'incendie et de sauvetage au monde.

L'effectif se compose en 2015 de 5 992 personnes dont 5 096 pompiers opérationnels (officiers compris), tous professionnels.

Andy Roe est l'actuel commissaire aux incendies de Londres (London Fire Commissioner, LFC). Il est responsable de la manière dont la London Fire Brigade fournit les services d'incendie et de secours à Londres.

La brigade des pompiers de Londres utilisent les différents matériels de secours ci-dessous :

- Des nacelles : camions équipés de grandes échelles
- Différentes échelles réparties stratégiquement dans toute la ville, ce qui leur permet de répondre aux urgences le plus rapidement possible.
- Des bateaux pour intervenir dans la Tamise notamment
- Des masques respiratoires
- D'une flotte de véhicules de secours adaptés.

Elle peut intervenir sur des incidents variés comme le secours à la personne, les accidents de la route, dans l'eau et les incendies.

Les incidents sont gérés différemment selon leur gravité.

Lors d'incidents majeurs, une structure de commandement est mise en place pour gérer les équipes et résoudre l'incident de la manière la plus sûre possible. Lorsqu'un incident prend de l'ampleur, généralement quatre véhicules de pompiers ou plus, une unité de commandement est envoyée sur les lieux de l'incident.

La LFB dispose de différents indicateurs KPI à atteindre, dont :

- Temps moyen d'arrivée du premier équipage (mensuel) : 6 min
- Temps moyen d'arrivée du second équipage (mensuel) : 8 min
- Premier équipage arrivé en moins de 10 min : 90 % des cas
- Premier équipage arrivé en moins de 12 min : 95 % des cas
- Nombre de fausses alarmes du fait d'alarmes de détection automatique de feux dans des édifices non résidentiel : 20 000 /an max
- Nombre de morts par cause de feux : 50 /an
- Nombre de blessés par cause de feux : 1000 /an
- Nombre de feux concernant des maisons : 1700 /an
- Nombre de feux concernant des appartements : 2400 /an
- Nombre de feux concernant des centres de soins : 330 /an

2. D'un point de vue économique

Toutes les décisions officielles concernant la London Fire Brigade sont approuvées par le London Fire Commissioner. Certaines décisions doivent également être approuvées par le maire ou le maire adjoint de Londres.

Il s'agit notamment de l'approbation du budget annuel de la brigade et du plan de gestion des risques communautaires (CRMP), qui définit le plan de la brigade pour la protection de Londres.

Le budget annuel de la brigade de pompiers de Londres se décide sur la base de la réalisation des indicateurs des années précédentes, en termes de réactivité et réussite des interventions.

La LFB accorde une grande attention aux indicateurs qu'elle soumet à la ville de Londres et qu'elle rend visible aux habitants.

Dans le plan 2023-2029 de mesure des performances de la LFB, présenté en mai 2022, on y retrouve trois types de mesures :

- les principales mesures de résultats, qui indiquent s'ils atteignent leurs objectifs à long terme;
- Les mesures clés des processus, qui indiquent s'ils apportent les améliorations nécessaires pour atteindre leurs objectifs à long terme ;
- Toute autre mesure présentant un intérêt significatif pour les communautés qu'ils servent.

II) Objectifs du projet

1. Principaux objectifs du projet

Les objectifs pour les équipes de la brigade sont les suivants:

- permettre d'aider les équipiers pompiers à améliorer leur temps de réponse aux incidents, c-a-d le temps d'arrivée sur les lieux,
- et d'avoir le bon nombre d'équipes déployées selon le type d'incidents et le lieu d'incidents.

Les objectifs pour la mairie de Londres et le commissaire aux incendies de Londres sont de réduire le coût des interventions futures (matériels, humains) pour les pompiers, et également de limiter le besoin de soins des victimes et de dégâts matériels causés par les incidents.

Cela dans un but global de maintien du plan annuel de la brigade pour la protection de Londres.

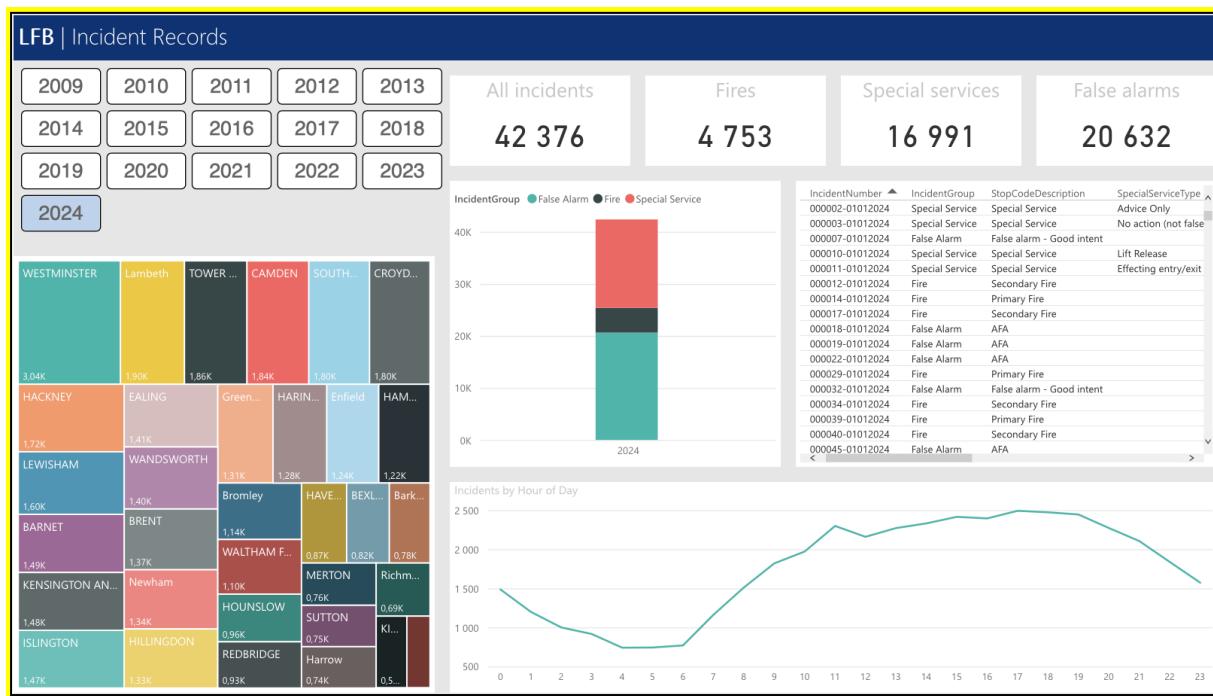


Figure 1 - Exemple de tableaux de bord de la LFB indiquant les incidents enregistrés

A notre niveau, notre objectif projet est de **prédire le temps de réponse à des interventions de la brigade des pompiers de Londres, fonction du nombre d'équipes déployées, du type d'incidents et du lieu de l'incident**.

2. Problématique

Comment prédire et optimiser le temps d'arrivée sur les lieux des incidents des pompiers de Londres en fonction des effectifs matériels, humains mis en œuvre, du type d'incidents, du lieu d'incidents et des données temporelles ?

3. Composition et maturité de l'équipe projet

Notre équipe projet fait partie de la promotion AVR24_CDA "Data Analyst".

Elle est composée de 4 personnes :

- Paul DEVIN, Consultant en recrutement
- Thomas GARRIGUE, Délégué du Numérique en Santé
- Marie HERBERT, Chef de projet informatique
- Namita KALA, Conseillère clientèle

Nos niveaux d'expertise autour de la problématique sont variés.

Paul : Aucune expertise mais un intérêt pour le sujet d'un point de vue technique comme pratique.

Thomas : Je n'ai pas d'expertise dans ce domaine mais j'ai un intérêt pour ce thème. Notamment le fait qu'il soit question d'interventions quotidiennes, sur différentes thématiques, qui doivent être rapides et efficientes pour limiter des risques et préserver des vies.

Marie : Je n'ai pas d'expertise spécifique sur la problématique adressée. J'ai une quelques connaissances du monde des pompiers et un intérêt pratique sur le sujet.

Namita : Je n'ai pas d'expertise dans ce domaine. J'ai essayé d'obtenir des informations sur internet.

4. Planning du projet - voir document Calendrier de projet-AVR24_CDA_Pompiers.pdf

PHASE	DÉTAILS	T2				T3				T4			
		MAI	JUIN	JUILLET	AOUT	SEPTEMBRE	OCTOBRE	NOVEMBRE	DÉCEMBRE				
1	SEMAINE DU PROJET : Cadrage et découverte du projet - Cadrage avec Raja (Mer. 29/05) - Découverte des données et du projet (Etape 1) - Coll'équipe projet 1 - Coll'équipe projet 2 - Rendu Rapport Etape 1 (Mar. 11/06)	6 13 20 27	3 10 17 24	1 8 15 22 29	5 12 19 26	2 9 16 23 30	7 14 21 28	4 11 18 25	2 9 16 23 30				
2	Exploration et analyse des données avec DataViz² - Cadrage avec Raja (Lun. 17/06) - Représentation graphique pertinente (Etape 2) - Coll'équipe projet 1 - Coll'équipe projet 2 - Coll'équipe projet 3 - Rendu Rapport Etape 2 (Ven. 28/06)			Etape 1 Coll 1 Coll 2	R Et 1 Coll 3								
3	Nettoyage et Pre-processing - Cadrage avec Raja (xx) - Nettoyage et pre-processing (Etape 3) - Coll'équipe projet 1 - Coll'équipe projet 2 - Coll'équipe projet 3 - Rendu Rapport Etape 3 (Ven. 19/07)			Etape 2 Coll 1 Coll 2	Etape 3 Coll 1 Coll 2 Coll 3	R Et 2 Coll 4	R Et 3 Coll 5						
4	Modélisation - Cadrage avec Raja (xx) - Élaboration des premières itérations de modélisation : (Step 1) - Coll'équipe projet 1 - Coll'équipe projet 2 - Coll'équipe projet 3 - Coll'équipe projet 4 - Coll'équipe projet 5 - Coll'équipe projet 6 - Rendu Rapport Etape 4 Step 1 (Ven. 30/08) - Interprétation, optimisation (Step 2) - Coll'équipe projet 7 - Coll'équipe projet 8 - Coll'équipe projet 9 - Coll'équipe projet 10 - Rendu Rapport Etape 4 (Ven. 27/09)			Etape 4 Step 1 Coll 1 Coll 2 Coll 3 Coll 4 Coll 5	Etape 4 Step 2 Coll 6 R Et 4 Step 1 Coll 7 Coll 8 Coll 9 Coll 10	R Et 4 Coll 11							
5	Clôture du projet - Cadrage avec Raja (xx) - Finalisation du code - Coll'équipe projet 1 - Coll'équipe projet 2 - Coll'équipe projet 3 - Coll'équipe projet 4 - Coll'équipe projet 5 - Coll'équipe projet 6 - Coll'équipe projet 7 - Rapport final (Ven. 25/10) - Préparation Soutenance finale - Coll'équipe projet 8 - Coll'équipe projet 9 - Coll'équipe projet 10 - Coll'équipe projet 11 - Coll'équipe projet 12 - Soutenance blanche - Soutenance finale (Semaine du 18/11)			Etape 5 Coll 1 Coll 2 Coll 3 Coll 4 Coll 5 Coll 6 Coll 7 Coll 8 Coll 9 Coll 10 Coll 11	Finalisation du code Coll 1 Coll 2 Coll 3 Coll 4 Coll 5 Coll 6 Coll 7 Coll 8 Coll 9 Coll 10 Coll 11 Coll 12	Soutenance finale Coll 1 Coll 2 Coll 3 Coll 4 Coll 5 Coll 6 Coll 7 Coll 8 Coll 9 Coll 10 Coll 11 Coll 12							

III) Cadre du projet

1. Jeux de données utilisés

a. Détail des incidents depuis 2009

Ce premier jeu de données contient les détails de chaque incident traité depuis janvier 2009. Des informations sont fournies sur la date et le lieu de l'incident ainsi que sur le type d'incident traité.

Nous pouvons retrouver tous les détails de ce jeu de données (nom des champs, description, type de champ, nombre de valeurs manquantes par champ et distribution des valeurs) dans les onglets “Incident” de l’Excel ci-joint :

- Incidents de 2009 à 2017 :

<https://docs.google.com/spreadsheets/d/1GqSMU8JqPfDn2ec-z6Bu-tv554g0mYxZ/edit#gid=1997183491>

Figure 2 - Liste des variables des jeux de données “Incidents”

Column	Sample record	Description
IncidentNumber	000008-01012018	LFB Incident Number
DateOfCall	01-Jan-18	Date of 999 call
CalYear	2018	Year of 999 call
TimeOfCall	00:04:25	Time of 999 call
HourOfCall	0	Hour of 999 call
IncidentGroup	False Alarm	High level incident category
StopCodeDescription	AFA	Detailed incident category
SpecialServiceType		Further detail for special services incident categories
PropertyCategory	Non Residential	High level property descriptor
PropertyType	Mosque	Detailed property descriptor
AddressQualifier	Within same building	Qualifies location of actual incident relevant to category above
Postcode_full	N2 8AY	Postcode
Postcode_district	N2	Postcode Districts
UPRN	200220110	Unique Property Reference Number
USRN	20013420	Unique Street Reference Number
IncGeo_BoroughCode	E09000003	Borough Code
IncGeo_BoroughName	BARNET	Borough Name
ProperCase	Barnet	Borough Name
IncGeo_WardCode	E05000049	Ward Code
IncGeo_WardName	EAST FINCHLEY	Ward Name
IncGeo_WardNameNew	EAST FINCHLEY	New Ward Name
Easting_m	527184	Easting
Northing_m	189488	Northing
Easting_rounded	527150	Easting rounded up to nearest 50
Northing_rounded	189450	Northing rounded up to nearest 50
Latitude	51.58990022	Latitude
Longitude	-0.165452578	Longitude
FRS	London	Fir Service ground
IncidentStationGround	Finchley	LFB Station ground
FirstPumpArriving_AttendanceTime	348	First Pump attendance time in seconds
FirstPumpArriving_DeployedFromStation	Finchley	First Pump deployed from station
SecondPumpArriving_AttendanceTime		Second Pump attendance time in seconds
SecondPumpArriving_DeployedFromStation		Second Pump deployed from station
NumStationsWithPumpsAttending	1	Number of stations with pumps in attendance
NumPumpsAttending	1	Number of pumps in attendance
PumpCount	1	
PumpHoursRoundUp	1	Time spent at incident by pumps, rounded up to nearest hour
Notional Cost (£)	328	Time spent multiplied by notional annual cost of a pump

- Incidents à partir de 2018 :
<https://docs.google.com/spreadsheets/d/1GqSMU8JqPfDn2ec-z6Bu-tv554g0mYxZ/edit#gid=160637744>

On peut dégager 4 grands axes à partir des colonnes de ces jeux de données :

1. Temporalité des incidents : Colonnes “DateOfCall” à “HourOfCall”
2. Types et propriétés des incidents : Colonnes “IncidentGroup” à “AddressQualifier”
3. Données géographiques des incidents : Colonnes “Postcode_full” à “IncidentStationGround”
4. Conditions et délais de prise en charge : Colonnes “FirstPumpArriving_AttendanceTime” à “PumpMinutesRounded”

b. Détail des interventions depuis 2009

Le second jeu de données contient les détails des interventions de chaque camion de pompiers envoyés sur les lieux d'un incident depuis janvier 2009.

Par camion de pompiers , on entend une équipe de pompiers, son nombre de moteurs (véhicules) et de pompes à incendie déployés.

Nous pouvons retrouver tous les détails de ce jeu de données (nom des champs, description, type de champ, nombre de valeurs manquantes par champ et distribution des valeurs) dans les onglets “Mobi_xxxx” de l'Excel ci-joint :

- Interventions de chaque camion de pompiers de 2009 à 2014 “Mobi_2009” :
<https://docs.google.com/spreadsheets/d/1GqSMU8JqPfDn2ec-z6Bu-tv554g0mYxZ/edit#gid=1911978334>
- Interventions de chaque camion de pompiers de 2015 à 2020 “Mobi_2015” :
<https://docs.google.com/spreadsheets/d/1GqSMU8JqPfDn2ec-z6Bu-tv554g0mYxZ/edit#gid=1844942983>
- Interventions de chaque camion de pompiers de 2021 à aujourd'hui “Mobi_2021” :
<https://docs.google.com/spreadsheets/d/1GqSMU8JqPfDn2ec-z6Bu-tv554g0mYxZ/edit#gid=47035469>

Column	Sample record	Description
IncidentNumber	000008-01012018	LFB Incident Number
CalYear	2018	Year of 999 call
HourOfCall	3	Hour of 999 call
ResourceMobilisationId	5055153	LFB Resource Mobilisation ID
Resource_Code	A392	LFB Resource Code
PerformanceReporting	1	First Pump arrived at incident
DateAndTimeMobilised	01/01/2018 00:04:25	Date and time of mobilised (GMT)
DateAndTimeMobile	01/01/2018 00:05:38	Date and time of mobile (GMT)
DateAndTimeArrived	01/01/2018 00:10:13	Date and time arrived (GMT)
TurnoutTimeSeconds	73	Turnout time in seconds
TravelTimeSeconds	275	Travel time in seconds
AttendanceTimeSeconds	348	Attendance time in seconds
DateAndTimeLeft	01/01/2018 00:16:38	Date and time left the incident (GMT)
DateAndTimeReturned	01/01/2018 00:22:45	Date and time returned to station (GMT)
DeployedFromStation_Code	A39	Deployed from station code
DeployedFromStation_Name	Finchley	Deployed from station name
DeployedFromLocation	Home Station	Deployed from location
PumpOrder	1	Pump order
PlusCode_Code	Initial	Code of Plus Code
PlusCode_Description	Initial Mobilisation	Description of Plus Code
DelayCodeId	9	Delay code ID
DelayCode_Description	Traffic, roadworks, etc	Delay code description

Figure 3 - Liste des variables des jeux de données "Interventions"

2. Accessibilité et volumétrie

a. Détail des incidents depuis 2009

Le 9 janvier 2014, dix casernes de pompiers londoniennes ont été fermées dans le cadre du cinquième plan de sécurité londonien (LSP5) de l'Autorité, et les zones des casernes de pompiers ont été modifiées pour refléter ces fermetures, les zones des casernes fermées étant réparties dans les zones des casernes de pompiers adjacentes.

Afin de fournir des données cohérentes sur les incidents, les noms des casernes ont été modifiés pour tous les incidents de cet ensemble de données et reflètent les zones des casernes utilisées depuis le 9 janvier 2014.

De cette information, on peut en déduire que nous attendons une répartition altérée des variables 14, 15 et 16 des fichiers "Mobi 2009", "Mobi 2015" et "Mobi 2021".

Accessibilité : Le jeu de données le plus récent est actualisé tous les trimestres pour le dataset “Incident”.

Volumétrie :

- LFB Incident Data from 2018 : **713 368 incidents**
- LFB Incident Data from 2009 to 2017 : **988 279 incidents**

Il y a donc **1 701 647 incidents** recensés depuis 2009 dans nos jeux de données à analyser.

Ces deux jeux de données comprennent le même nombre de colonnes, à savoir 39.

b. Détail des interventions depuis 2009

Le détail des interventions de la London Fire Brigade est disponible via 3 datasets. Ils contiennent le détail de chaque camion de pompier envoyé sur un incident depuis janvier 2009.

Accessibilité : Le jeu de données le plus récent est actualisé tous les mois pour le dataset “Mobilisation”.

Volumétrie :

- LFB Mobilisation data from 2021 - 2024 : **587 919 interventions**
- LFB Mobilisation data from 2015 - 2020 : **883 641 interventions**
- LFB Mobilisation data from 2009 - 2014 : **901 788 interventions**

Il y a donc **2 373 348 interventions** recensées depuis 2009 dans nos jeux de données à analyser.

Ces trois jeux de données comprennent le même nombre de colonnes, à savoir 22.

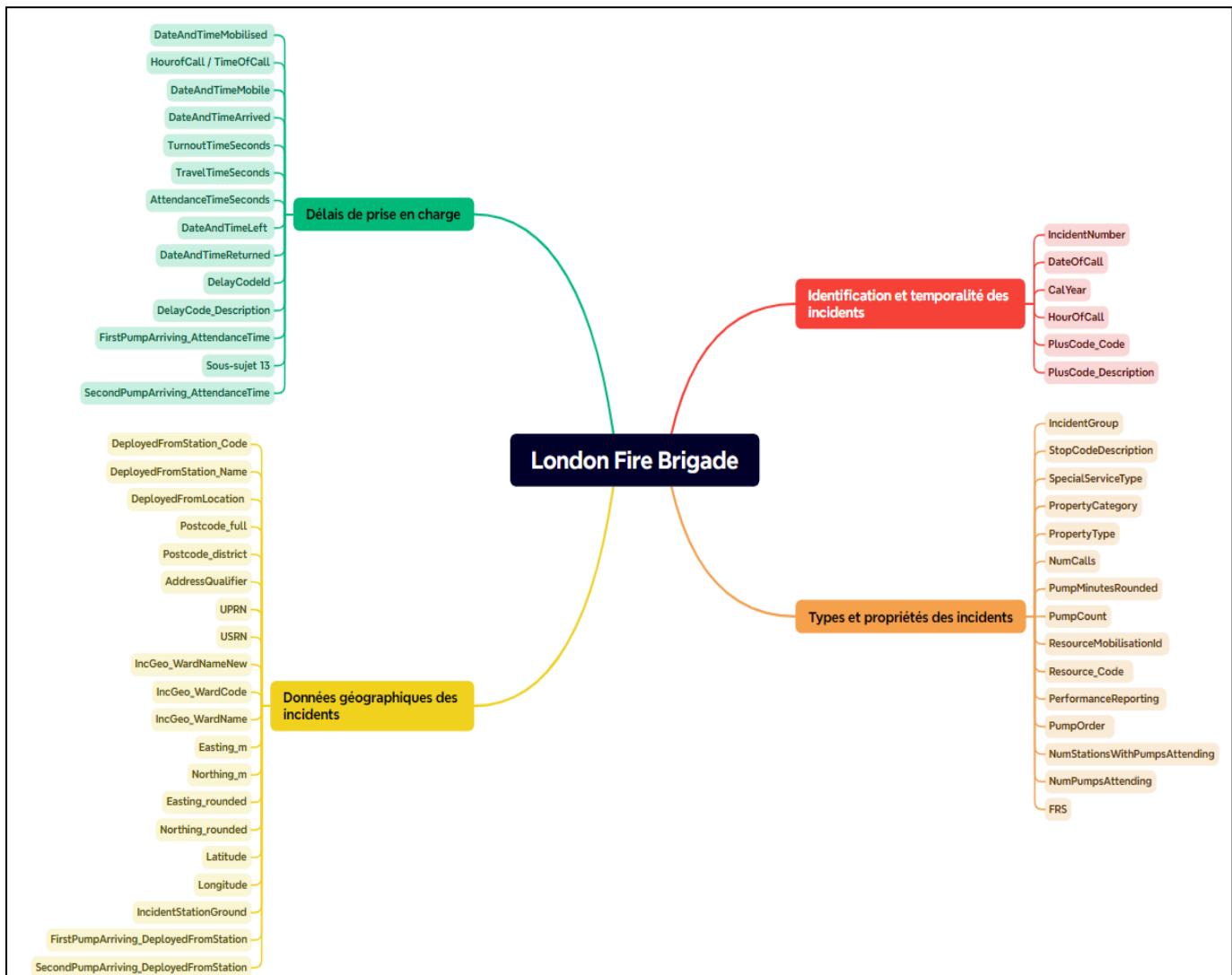
IV) Zoom sur les données

1. Pertinence des variables

Nous pouvons regrouper nos variables en catégories que nous emploierons à l'élaboration de nos modèles de machine learning, nous pouvons ainsi différencier les catégories de variables détaillées plus haut:

- Identification et temporalité des incidents : Y'a -t'il une temporalité propice au déclenchement de chaque type d'incident?
- Types et propriétés des incidents: En quoi le type d'accident influe-t'il sur le temps d'intervention?
- Données géographiques des incidents : La caserne sollicitée est-elle la plus proche du lieu d'incident? Certaines zones géographiques sont-elles plus sujettes aux incidents et si oui de quel type?
- Délais de prise en charge: Nous permet de séquencer les temps d'interventions, pour mettre en lumière les facteurs qui les influencent pour mieux les prédire et les améliorer.

Figure 4 - Mapping des catégories de variables pertinentes pour notre modèle



2. Variable cible

Segmentation du temps d'intervention :

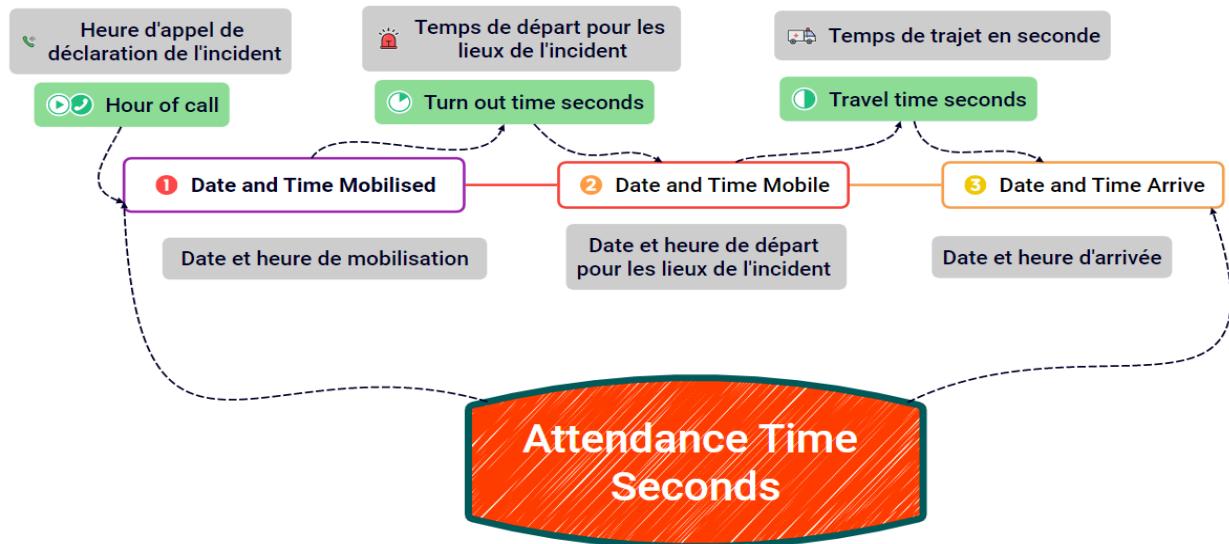


Figure 5 - Segmentation du temps d'intervention

La variable cible que nous allons chercher à prédire est le temps d'arrivée sur les lieux de l'incident (AttendanceTimeSeconds).

3. Particularités remarquables du jeu de donnée

Dans le cadre d'une première exploration nous avons cherché à déterminer les corrélations entre nos variables avant jointure des fichiers.

Sur les heatmap réalisées ci-dessous, on constate que la majorité des variables n'ont pas de relation de corrélation, si ce n'est les corrélations évidentes entre AttendanceTimeSeconds et TravelTime Second, AttendanceTimeSeconds et PumpOrder et DelayCoded.

Nous verrons dans un second temps les corrélations existantes entre les variables une fois les 5 dataset joints.

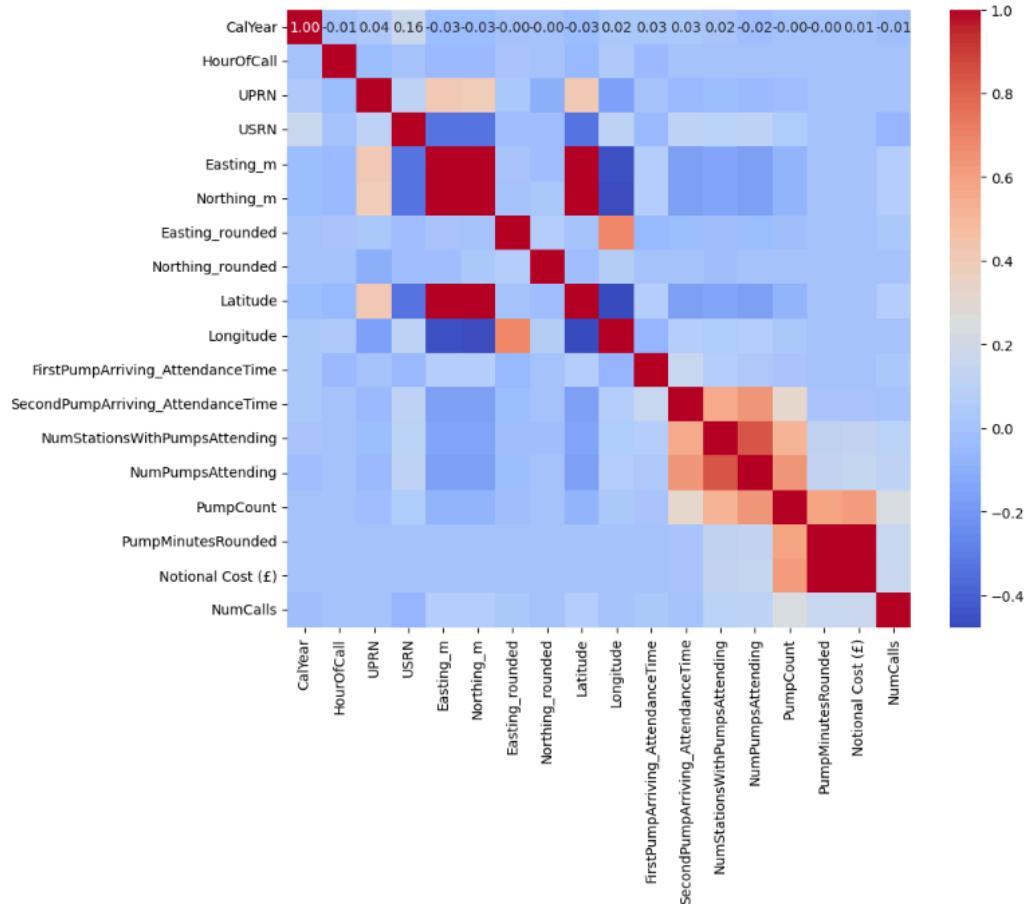


Figure 6 - Heatmap des incidents

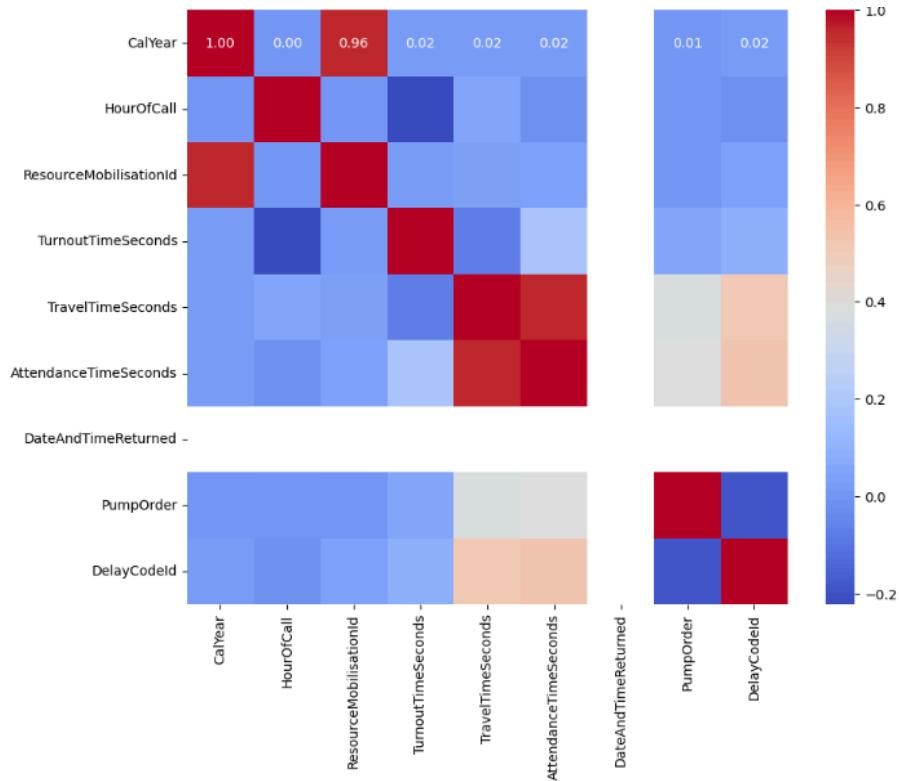


Figure 7 - Heatmap des mobilisations/interventions

4. Données limitantes

Première donnée

Afin de préciser l'étude de nos données, il serait intéressant de croiser les temps d'intervention et de trajet au trafic de Londres à certaines heures de la journée, pour identifier la perturbation de la circulation.

Nous réfléchirons à peut-être utiliser cette carte de la ville de Londres avec le trafic en temps réel : <https://tfl.gov.uk/info-for/open-data-users/api-documentation>

Deuxième donnée

Nous ne connaissons pas l'application de navigation utilisée par les équipes pour se rendre sur les lieux des interventions.

Selon l'application utilisée, le temps de trajet pourrait être réduit.

Troisième donnée

Nous n'avons pas d'informations très détaillées sur le type d'incidents et d'interventions hormis : 'Special Service', 'Fire', 'False Alarm'.

Nous pourrions utiliser la variable type de service spécial "SpecialServiceType" pour identifier le type d'interventions réalisé comme :

'RTC', 'Assist other agencies', 'Flooding', 'Medical Incident', 'Lift Release', 'No action (not false alarm)', 'Effecting entry/exit', 'Animal assistance incidents', 'Other rescue/release of persons', 'Making Safe (not RTC)', 'Other Transport incident', 'Evacuation (no fire)', 'Spills and Leaks (not RTC)', 'Removal of objects from people', 'Hazardous Materials incident', 'Advice Only', 'Water provision', 'Rescue or evacuation from water', 'Suicide/attempts', 'Stand By', 'Medical Incident - Co-responder'.

Cependant, cette variable comporte environ 60% de valeurs manquantes. En effet, les 40% représentent peut-être des interventions spécifiques où une brigade spéciale a été mobilisée.

V) Exploration et analyse des données avec DataViz'

1. Relation entre les variables explicatives et cibles

Comme nous l'avons vu précédemment, la variable cible est "AttendanceTimeSeconds". Le but étant de prédire et d'estimer le temps d'arrivée sur les lieux d'incidents des brigades. Nous appellerons dans la suite de notre développement cette valeur "Temps d'intervention".

Les variables explicatives sont les variables qui influencent notre variable cible.

Nous retrouverons ainsi :

- 'HourOfCall', 'DateAndTimeMobilised', 'DateAndTimeMobile',
'DateAndTimeArrived', 'DateAndTimeLeft', 'DeployedFromStation_Code',
'DeployedFromStation_Name', 'DeployedFromLocation', 'PumpOrder',
'PlusCode_Code', 'PlusCode_Description', 'DelayCodeId',
'DelayCode_Description', 'IncidentGroup', 'StopCodeDescription',
'SpecialServiceType', 'PropertyCategory', '.PropertyType', 'AddressQualifier',
'Postcode_full', 'Postcode_district', 'UPRN', 'USRN', 'IncGeo_BoroughCode',
'IncGeo_BoroughName', 'IncGeo_WardCode', 'IncGeo_WardName',
'IncGeo_WardNameNew', 'Latitude', 'Longitude', 'FRS',
'IncidentStationGround', 'FirstPumpArriving_AttendanceTime',
'FirstPumpArriving_DeployedFromStation',
'SecondPumpArriving_AttendanceTime',
'SecondPumpArriving_DeployedFromStation',
'NumStationsWithPumpsAttending', 'NumPumpsAttending', 'PumpCount',
'PumpMinutesRounded', 'Notional Cost', 'NumCalls'.

Nous avons également pu identifier dans le mapping réalisé en Figure 4 (Mapping des catégories de variables pertinentes pour notre modèle), la relation entre les variables explicatives.

2. Distributions des données

a. Brainstorming de la Dataviz

Afin d'identifier de façon macro l'influence des variables explicatives sur la variable cible, avant de nettoyer le jeu de données, nous avons brainstormé pour établir une liste de quelques data visualisation pertinentes à réaliser.

Voici les résultats de notre brainstorming :

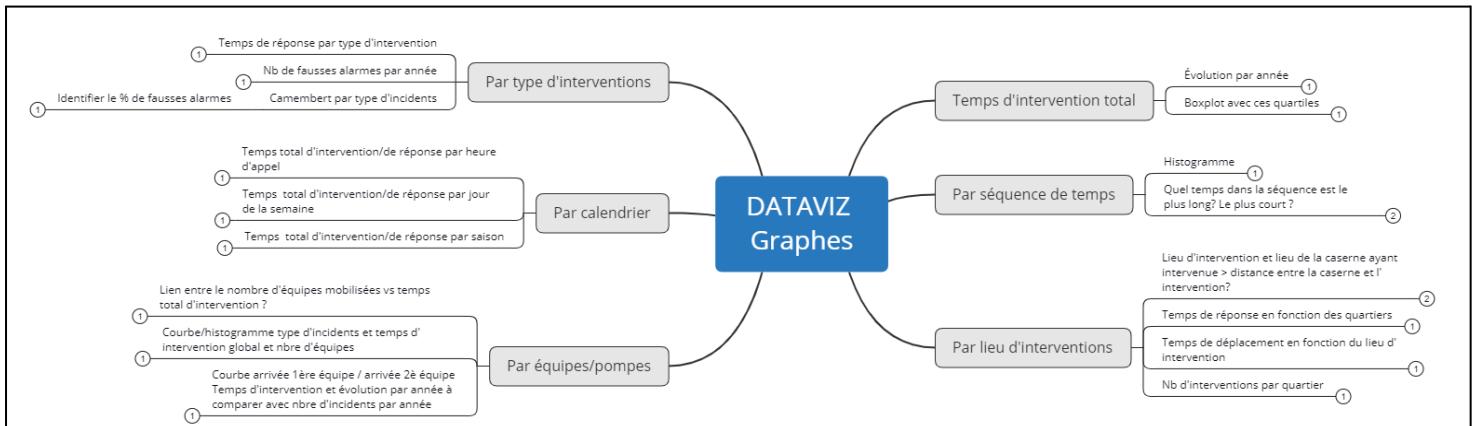


Figure 8 - Brainstorming de la data visualisation

b. Data visualisation

Pour réaliser cette data visualisation, nous avons mergé les datasets "Mobilisations" et "Incidents" afin de pouvoir croiser les jeux de données et identifier les relations entre les variables des 2 datasets.
Le dataframe créé est nommé "df".

i. Analyse du temps d'intervention

Première visualisation

Construction: Pour réaliser cette visualisation, nous avons utilisé la méthode barplot de la bibliothèque Seaborn.

La variable 'x' est l'année de l'appel 'CalYear', la variable 'y' est le temps d'intervention total en secondes, selon le type d'incident "Incident Group" placé en argument 'hue'. Tout cela basé sur le jeu de données "df".

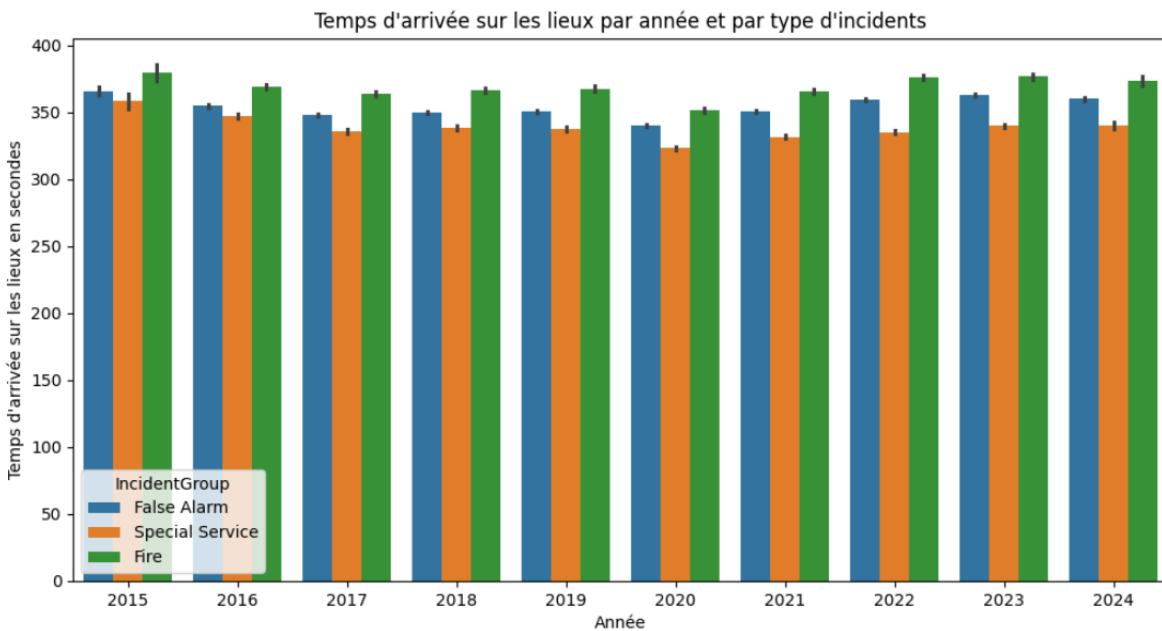


Figure 9 - Evolution du temps d'arrivée sur les lieux par année et par type d'incidents

Analyse :

Nous observons que les données sont calculées de l'année 2015 à 2024.

Les temps d'interventions totaux oscillent entre 325 et 375 secondes.

La moyenne de ce temps d'intervention est quasi constante et la barre d'erreur est très faible et indépendante de l'année et du type d'incident.

Il est à noter que les False Alarm représente un temps d'intervention semblable à ceux des "vraies" alarmes de type Spécial Service et Fire.

Ceci est problématique car certaines casernes peuvent perdre du temps à intervenir sur des fausses alarmes et nécessitent l'intervention d'une caserne plus lointaine sur un autre vrai incident ayant eu lieu au même moment, ce qui pourrait allonger le temps d'intervention total..

Nous décrirons dans un prochain graphe la répartition des "FalseAlarm".

Seconde visualisation

Construction : Pour réaliser cette visualisation, nous avons utilisé la méthode boxplot de la bibliothèque Seaborn.

La variable 'x' est l'année de l'appel 'CalYear', la variable 'y' est le temps d'intervention total en secondes, selon le type d'incident "Incident Group" placé en argument 'hue'. Tout cela basé sur le jeu de données "df".

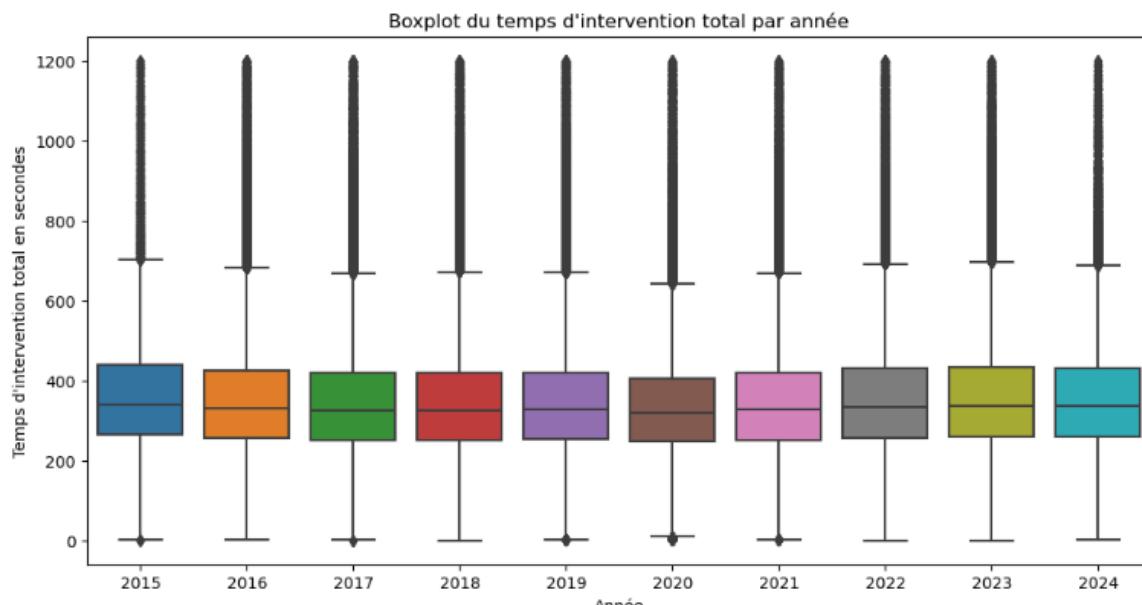


Figure 10 - Boxplot du temps d'intervention (= arrivée sur les lieux) par année

Analyse :

Ce graphe confirme le précédent selon lequel le temps d'intervention n'a pas de lien avec l'année de l'incident.

En effet, le temps d'intervention est très centré autour de 350 secondes environ.

Le boxplot a des quartiles entre 250 secondes et 425 secondes.

Nous retrouvons des outliers entre 700 et 1200 secondes, qui sont eux aussi semblables chaque année.

ii. Analyse de la séquence de temps

Première visualisation

Construction : Pour réaliser cette visualisation, nous avons utilisé l'histogramme de la bibliothèque seaborn. Cette figure comporte 3 histogrammes, représentant le séquencement des temps que nous avons à disposition. Le dernier représentant le temps d'intervention.

Analyse : On remarque que les 3 graphiques représentent des données suivant une distribution à queue lourde, avec une moyenne de temps total de réponse avoisinant les 350, 375 secondes. On note néanmoins un nombre important de valeurs maximum cumulées.

Comme vu précédemment le temps de mobilisation et la durée la plus faible des temps.

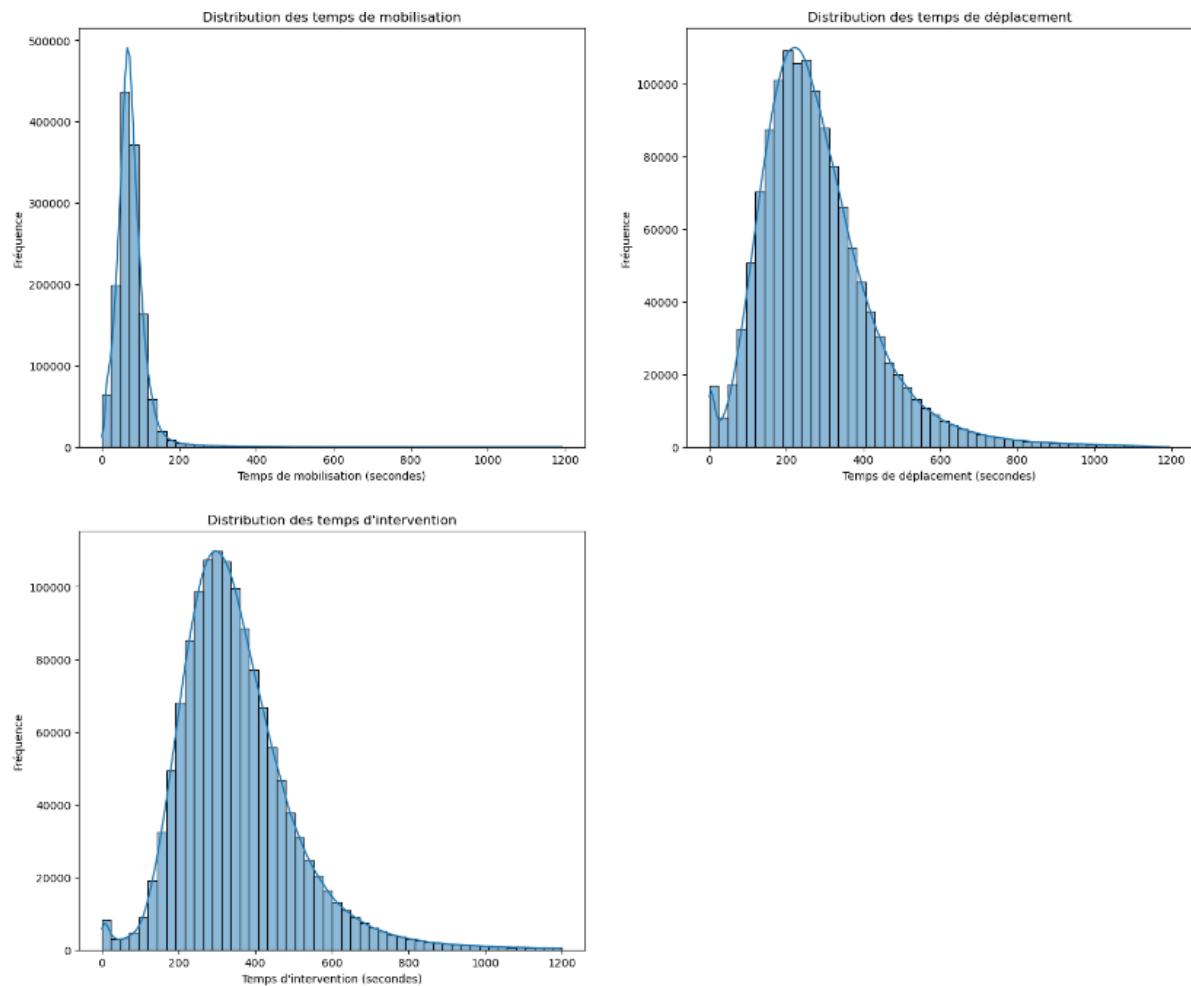


Figure 11 - Distribution des séquences de temps

Seconde visualisation

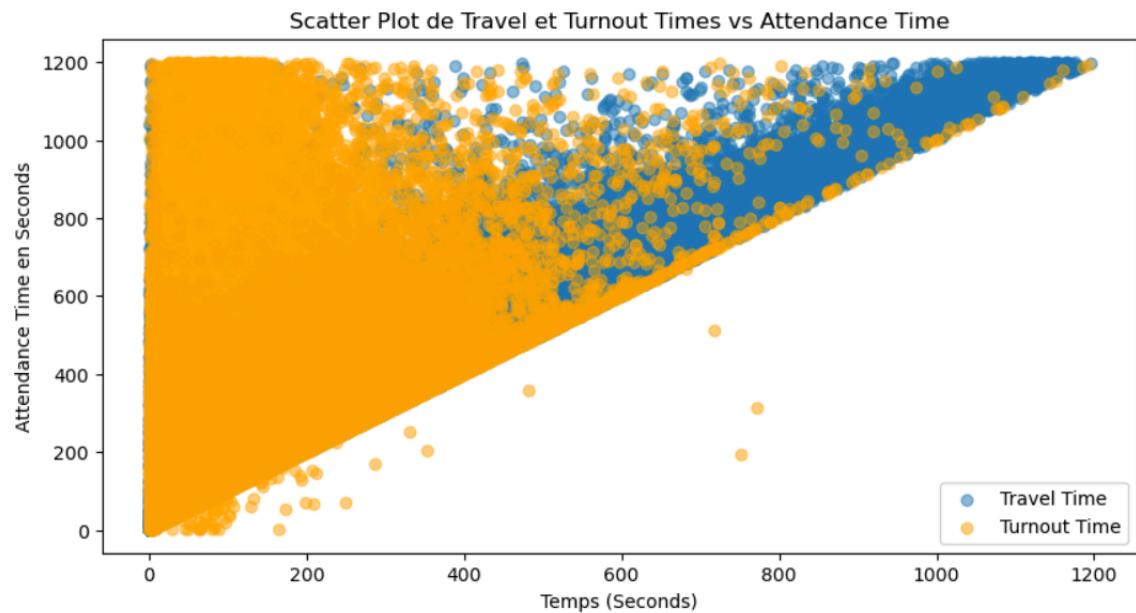
Construction : Pour réaliser cette visualisation, nous avons utilisé un graphique en nuage de points qui montre la relation entre TravelTimeSeconds, TurnoutTimeSeconds et AttendanceTimeSeconds (temps de déplacement, temps de mobilisation et temps d'arrivée).

Analyse : On remarque que les points bleus (TravelTimeSeconds) montrent une corrélation positive avec AttendanceTimeSeconds. À mesure que le temps de déplacement augmente, le temps de réponse total augmente également. Cela indique que le TravelTimeSeconds est un composant significatif du temps de réponse total.

Les points orange (TurnoutTimeSeconds) montrent également une corrélation positive avec AttendanceTimeSeconds. Cependant, la distribution est plus dense à

l'extrémité inférieure de l'axe des X, ce qui indique que les TurnoutTime sont généralement plus courts par rapport aux Traveltime.

La dispersion des points bleus est plus large sur l'axe des X par rapport aux points orange . Cela indique que le TravelTime a une gamme de valeurs plus large et potentiellement un impact plus grand sur le temps de réponse total.



*Figure 12 - Corrélation entre TravelTime, TurnoutTime et AttendanceTime
(temps de déplacement, temps de mobilisation et temps d'arrivée)*

iii. Analyse du lieu d'interventions

Première visualisation

Construction : Pour réaliser cette visualisation, nous avons utilisé la méthode countplot de la bibliothèque Seaborn.

La variable 'y' analysée est le nom de l'arrondissement, basé sur le jeu de données "df". Nous avons supprimé les doublons sur la variable IncidentNumber.

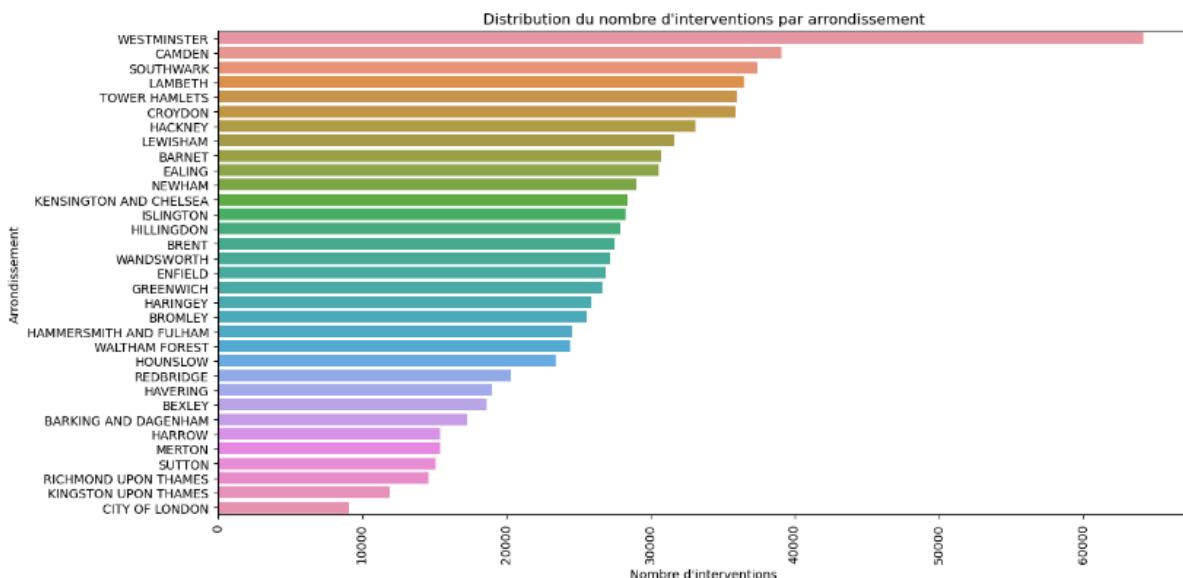


Figure 13 - Distribution du nombre d'interventions par arrondissement

Analyse :

Nous pouvons observer que le plus grand nombre d'interventions a lieu dans le quartier de Westminster, avec plus de 60 000 interventions sur l'ensemble du dataset. Ces interventions représentent 7.6% des interventions réalisées.

Viennent derrière les arrondissements de Camden, Tower Hamlets et SouthWark, pour plus de 35 000 interventions, soit 4.66%.

→ L'arrondissement est donc une variable explicative importante à conserver.

Seconde visualisation

Construction : Pour réaliser cette heatmap, Nous avons arrondi à 2 décimales la Latitude et la Longitude pour obtenir des zones géographiques plus larges simplifiant l'analyse. Nous avons ensuite calculé la moyenne des temps d'intervention pour chaque zone géographique. Nous avons enfin créé une table pivot pour préparer les données à être visualisée sous forme de carte. La heatmap provient de la librairie seaborn.

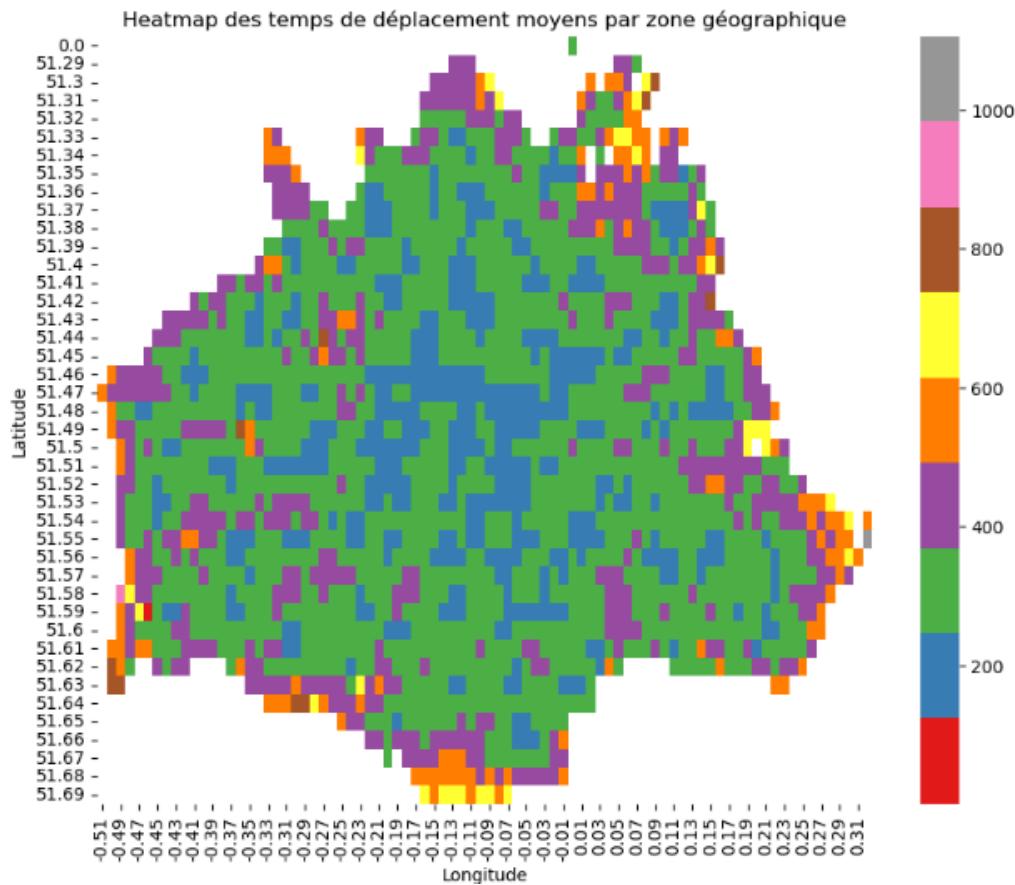


Figure 14 - Temps d'intervention moyen par zone géographique (Londres)

Analyse : Nous pouvons voir sur cette carte que les temps d'intervention moyen sont plus courts en centre-ville que sur la périphérie, nous dépassons les 1000 secondes sur certains secteurs. Une piste intéressante de réflexion serait de comprendre d'où partent les véhicules et en fonction du temps de trajet effectué et d'établir si le véhicule est parti de la station la plus proche.

Troisième visualisation

Construction : Pour construire notre carte, nous avons utilisé le package folium qui nous permet de créer une carte interactive en se basant sur les coordonnées géographiques des stations et des incidents répertoriés dans notre jeu de données. Nous pouvons ainsi visualiser le nombre de pompes envoyées par incidents par arrondissement, ainsi que la distance de l'incident par rapport à la station la plus proche.

Analyse : Plus la couleur est foncée, plus il y a d'incidents dans le quartier. Plus on va en périphérie de Londres, moins il y a d'incidents. En hyper-centre de Londres,

sont répertoriés près de 30 000 incidents de 2021 à mi-2024. Nous pouvons observer un triangle de 3 stations qui concentrent le plus d'incidents pour ces stations.

Nous pourrions préconiser d'augmenter le nombre de stations dans ce secteur.

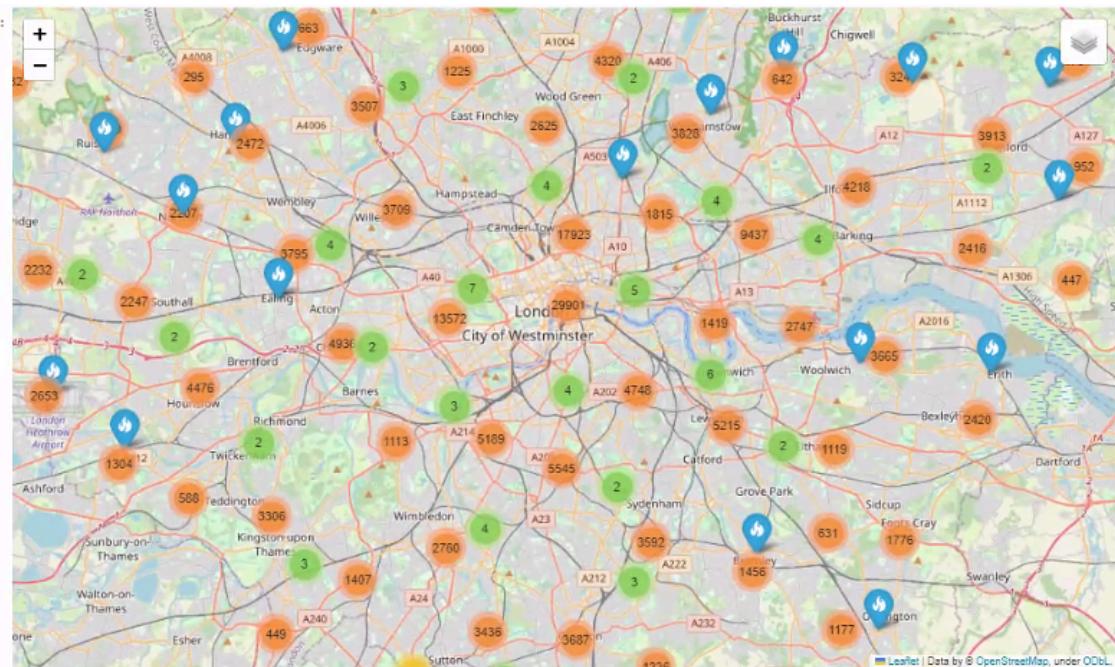


Figure 15 - Carte interactive du nombre de pompes déployées des incidents

iv. Analyse du type d'interventions

Première visualisation

Construction: Ce graphique a pour but de déterminer le nombre de 'False Alarm' par année. Pour cela il est intéressant de réaliser ce graphique en gardant les 2 autres types d'incidents et ainsi comparer quantitativement les évolutions des 3 groupes au fil des ans.

Nous avons ajouté les volumes sur les graphiques pour avoir une vision du nombre de fausses alarmes chaque année.

Pour créer ce graphique nous avons utilisé la bibliothèque seaborn.

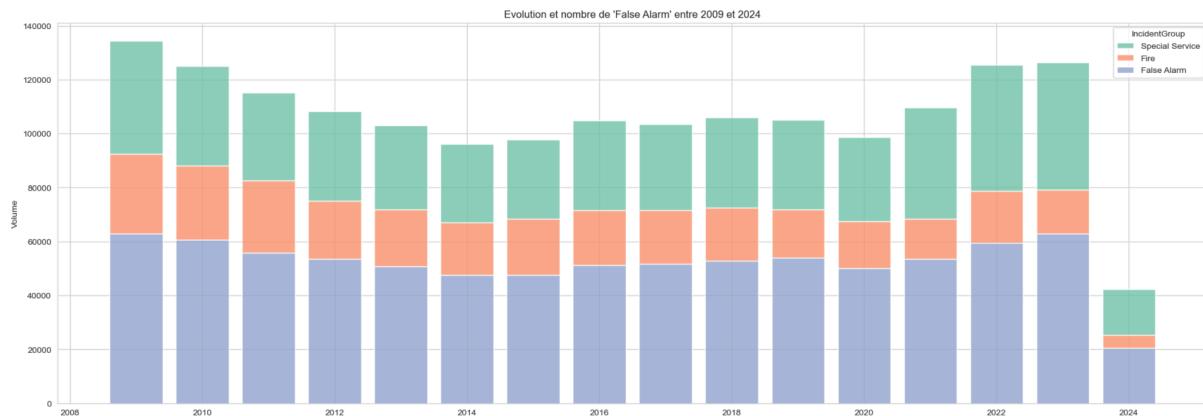


Figure 16 - Evolution et volume des fausses alarmes par année

Analyse : Les fausses alarmes étant représentées avec la couleur bleue, il s'agit donc du type d'incident le plus représenté chaque année. Au niveau des volumes, il y a des disparités en fonction des années mais les fausses alarmes semblent représenter environ la moitié des incidents chaque année.

Seconde visualisation

Construction : Le but de ce graphique est d'identifier le pourcentage de fausses alarmes dans le volume global des incidents ("incidentgroup"). Nous avons effectué un décompte du nombre total des 3 types d'incidents et réalisé un pie chart à l'aide de plotly.

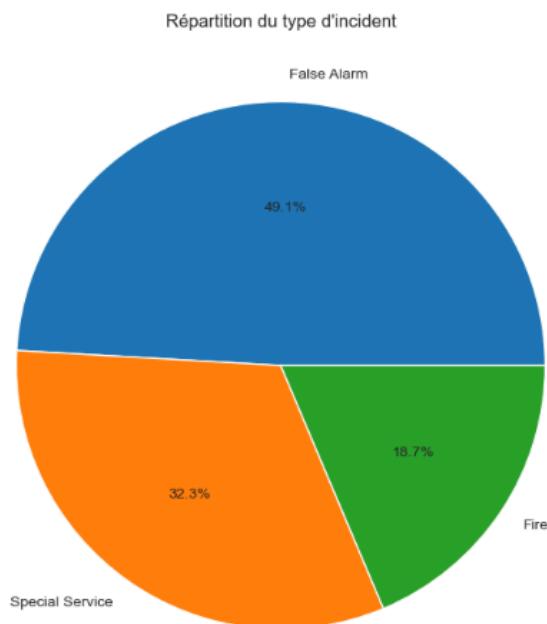


Figure 17 - Répartition des types d'incidents

Analyse : On constate que le nombre de fausses alarmes dans le volume total des incidents est très élevé. Il représente près de 50 % des incidents totaux entre 2009 et 2024. Cela confirme donc la vision que pouvait donner le graphique précédent.

Troisième visualisation

Construction : Si l'on va dans le détail des incidents avec la description secondaire issue des 3 catégories principales, cela nous permet de mettre en avant le type de fausses alarmes.

Pour cela nous avons réalisé un “sunburst” sous plotly.

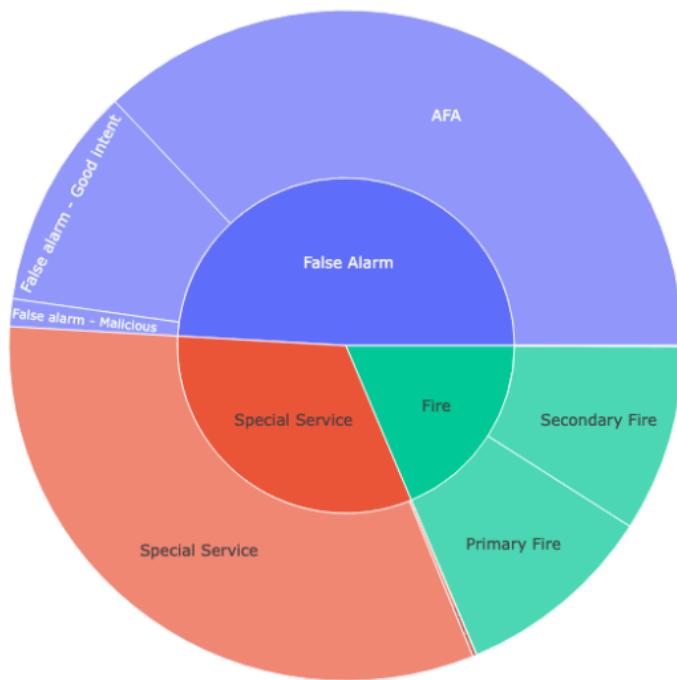


Figure 18 - Catégories détaillées des incidents

Analyse : On constate donc que :

- Les AFA ou “Automatic Fire Alarm” représente la majorité des cas
- Les Fausses alarmes de “bonne intention” représentent le second grand groupe
- Finalement les fausses alarmes “malveillantes” ne représentent qu'un volume très faible du total des fausses alarmes.

Quatrième visualisation

Construction : Le but de ce pie chart sous plotly est de détailler le type d'incident "Special Service" qui comme on le voit plus haut n'est pas sous catégorisé. Il existe toutefois une variable "SpecialServiceType" qui détaille le type d'action lorsque un tel incident se produit. Potentiellement certaines actions prennent plus de temps et influent donc sur le temps d'intervention final.

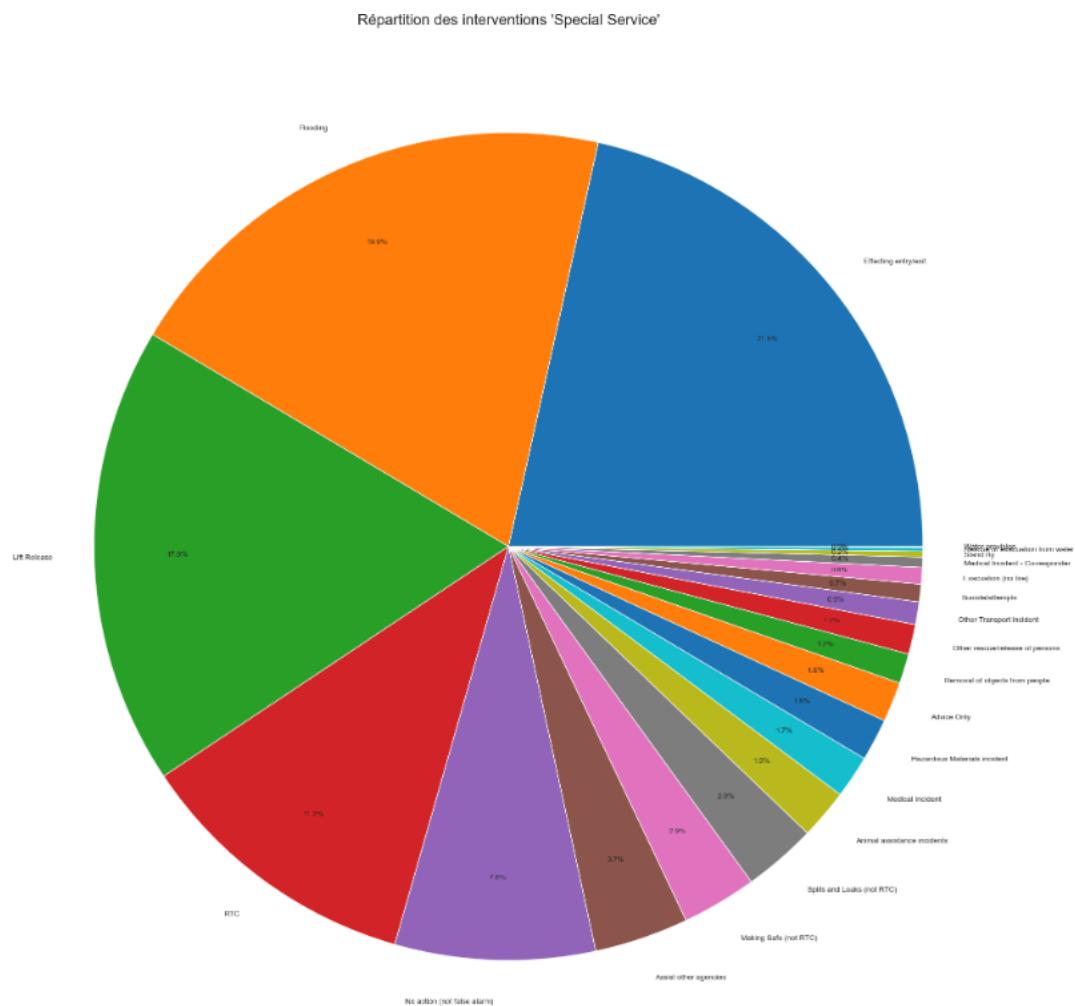


Figure 19 - Focus sur les actions issues des incidents de type 'special service'

Analyse : On constate bien ici que sur des interventions autres que les feux, les pompiers de Londres interviennent pour réaliser des actions multiples.

Dont les 4 plus importantes sont :

- Les évacuations
- Les inondations
- Les interventions sur ascenseur
- Les Road Traffic Collision (RTC)

v. Analyse du volume d'interventions par année, mois, jour, heure

Première visualisation

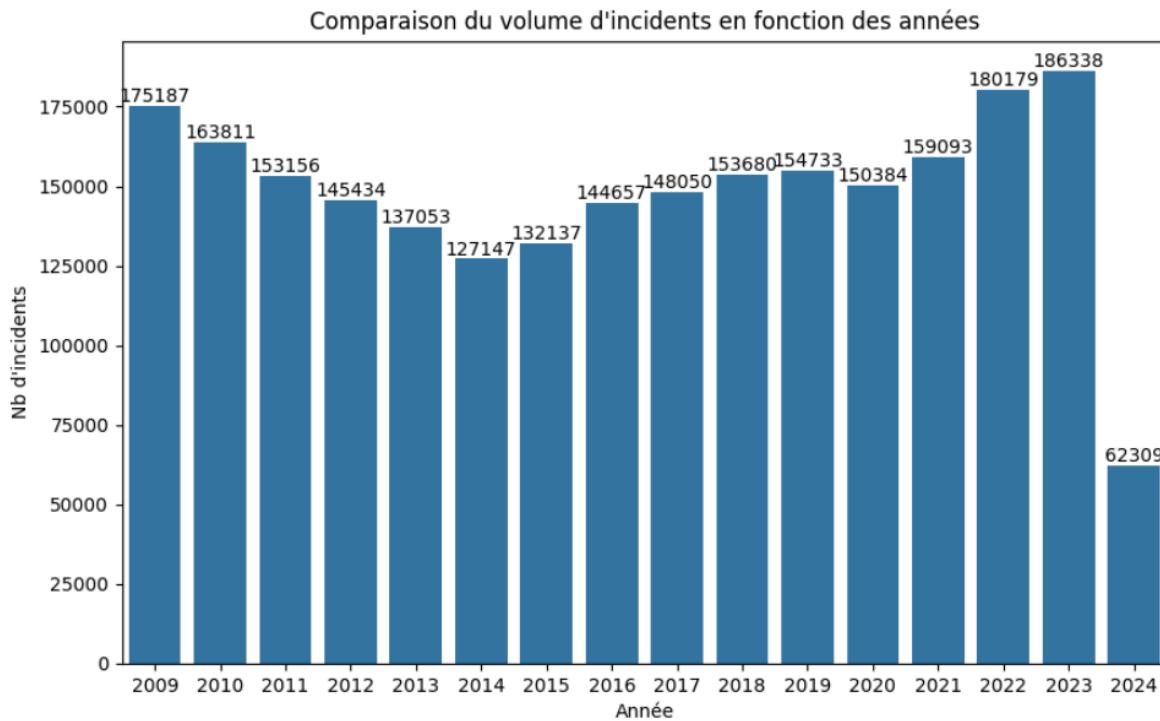


Figure 20 - Volume d'incidents en fonction des années (dataset = mobilisations)

Construction :

Ce barplot nous montre le volume d'incidents déclarés par année.

Nous nous sommes basées sur le dataset “mobilisations” car il contient des numéros d'incidents uniques, alors que le dataset “df” mergé duplique les lignes d'incidents qui ont le même “IncidentNumber”, car plusieurs équipes avaient pu intervenir.

Pour ce graphe, nous réalisons un “groupby” sur l'année extraite de la variable “DateandTimeMobilised” et un count() sur l'identifiant de l'incident “IncidentNumber”.

Analyse :

Entre 2009 et 2014, il y a eu une chute d'incidents de 175 000 à 127 000 incidents déclarés par an. Peut-être dû à de la prévention faite par les brigades de pompiers pour éviter de faire des faux appels ou car moins de personnes avaient des besoins.

Puis, il y a une augmentation progressive des incidents à partir de 2015. Avec un palier en 2018. Une légère baisse en 2020, que l'on peut sûrement lié au COVID.

Il semble qu'à partir de 2022, le nombre d'incidents augmentent largement de + 20 000 incidents par an, avec un pic sur l'année 2023 de plus de 185 000 incidents, alors que les années de 2016 à 2021 avaient un nombre d'incidents entre 144 000 et 158 000 incidents par an.

Plusieurs hypothèses : soit le dataset n'a pas toutes les données des années 2012 à 2016, bien que cela semble étonnant puisque l'ordre de grandeur du nombre d'incidents est constant, soit il y a eu une différence de méthode de saisie des incidents à partir de 2022., soit il y a eu réellement plus d'incidents à partir de 2022 et moins entre 2010 et 2015.

Deuxième visualisation

Construction :

Ce barplot présente le nombre d'interventions par heure de la journée.

Il a été réalisé sur le dataframe "df" mergé en faisant un "groupby" sur l'heure extraite de la variable "DateandTimeMobilised" et un count() sur l'identifiant de l'incident "IncidentNumber".

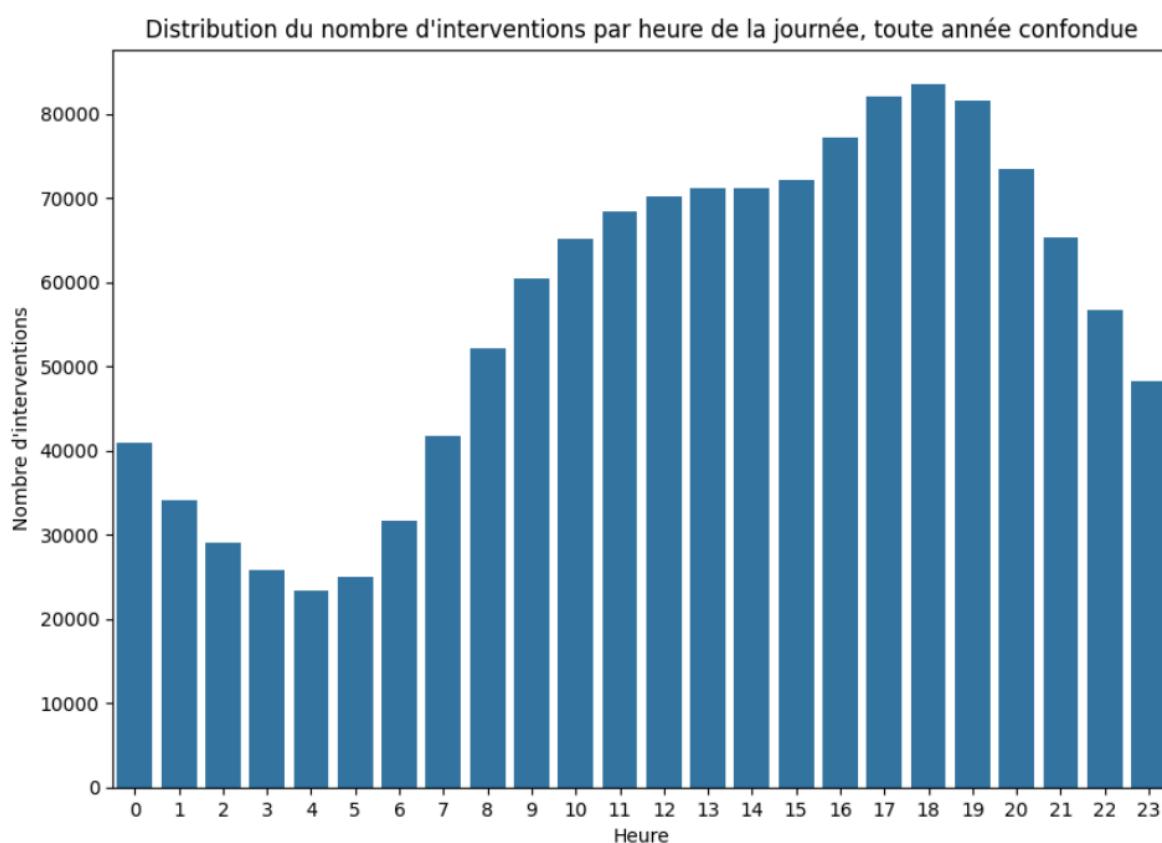


Figure 21 - Volume d'interventions par heure de la journée

Analyse : En toute logique, il y a moins d'interventions la nuit entre minuit et 6h du matin. Le nombre d'interventions commence à remonter progressivement à

partir de 7h du matin et se stabilise autour de 70 000 interventions, recensées entre 2015 et 2024, entre 11h et 15h. Il augmente avec un palier de + 5 000 à partir de 16h et jusqu'à 19h, heure de sortie du travail, des écoles où le trafic est plus dense et où les activités du quotidien reprennent. Il atteint un pic à 18h. Puis rechute de nouveau en soirée à partir de 20h.

Analyse complémentaire

D'autre part, nous n'avons pas noté de tendance notable de volume d'interventions fonction des jours de la semaine, ni par mois et par saison.

vi. Analyse du nombre d'équipes et de pompes

Première visualisation

Construction :

Ce nuage de points nous montre les temps d'intervention par nombre de pompes déployées et en fonction de la station d'où partent les équipages. Nous avons utilisé les variables suivantes: x="NumPumpsAttending", y="TotalResponseTime", hue="IncGeo_BoroughName".

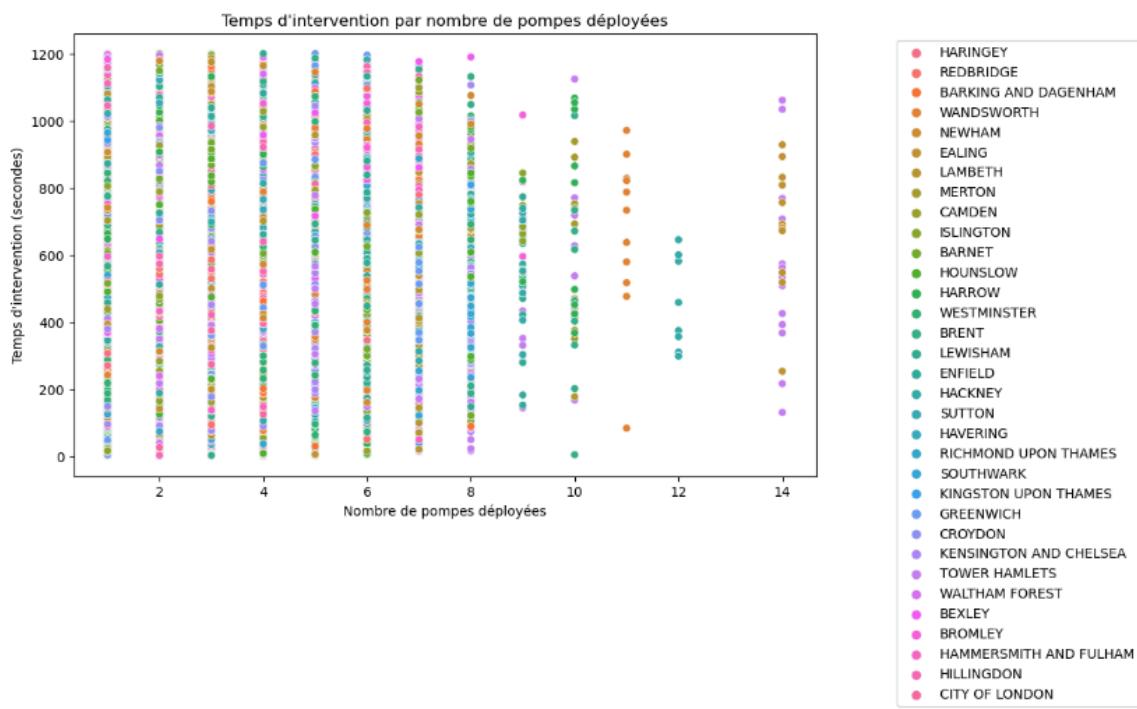


Figure 22 - Temps d'intervention par nombre de pompes déployées

Analyse : Il est difficile de voir une corrélation entre une évolution du temps d'intervention en rapport avec le nombre de pompes déployées, les deux

variables ne semblent pas corrélées. Nous pourrons en revanche pousser l'analyse sur les valeurs extrêmes pour tenter d'identifier des cas particuliers.

Seconde visualisation

Construction : L'idée de ce graphique est de visualiser le temps d'intervention de la 1ère équipe et aussi de la 2ème équipe. En effet, la LFB est objectivé avec un temps moyen d'intervention à 6 min de la 1ère équipe et 8 min pour la 2ème. Il a donc été effectué un 'groupby' en fonction des années associé aux variables "FirstPumpArriving_AttendanceTime" et "SecondPumpArriving_AttendanceTime". Nous avons également ajouté deux lignes de références horizontales indiquant le temps moyen d'intervention de la première et deuxième équipe.

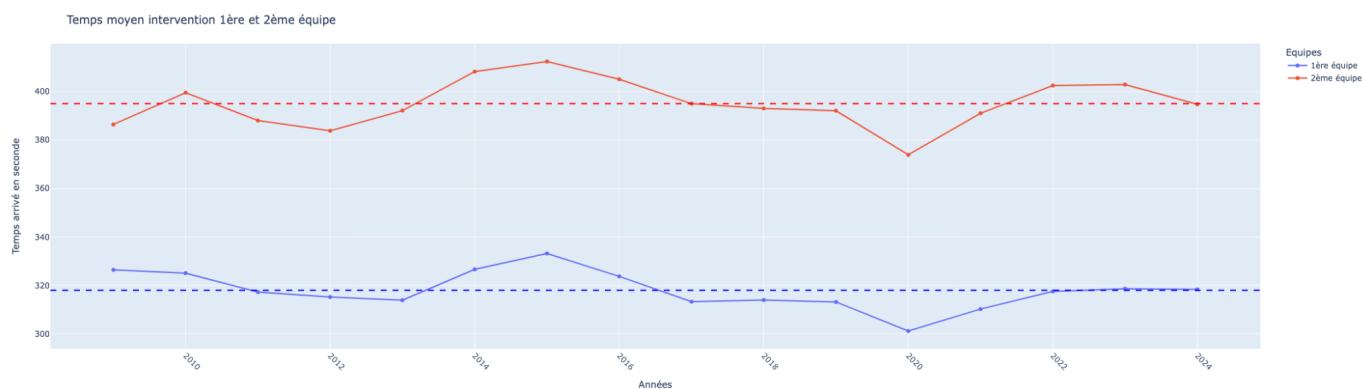


Figure 23 - Temps moyen d'intervention 1ère et 2ème équipe

Analyse : Le temps d'intervention de la première et de la deuxième équipe varie dans le temps. il semble d'ailleurs y avoir une relation dans le sens où lorsque sur une année le temps de la première équipe augmente cela sera le cas pour la deuxième équipe.

Toutefois, les équipes de la LFB n'ont jamais dépassé les valeurs cible de 6 min (360 secondes) pour la première équipe et de 8 min (480 secondes) pour la deuxième équipe.

vii. Analyse des temps de retards

Première visualisation

Construction : Pour construire les deux graphiques suivants nous avons conservé les valeurs non nulles du data frame, puis compté les itérations de ses valeurs. Les 2 graphiques sont des barplot issus de la bibliothèque seaborn.

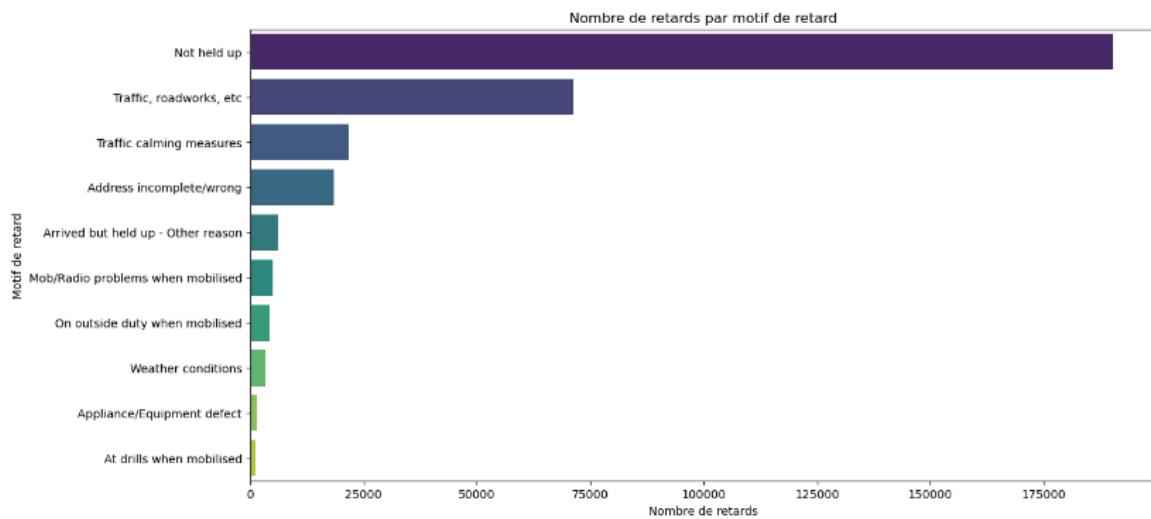


Figure 24 - Nombre de retards par motif de retard

Analyse : Sur ce graphique nous voyons la répartition du nombre de retard triés par motifs. La variable revenant principalement est "Not held up" que nous traduisons par non prise en charge. Ce motif n'étant pas quelque chose sur lequel on peut agir pour réduire le délai d'intervention, nous présentons dans un deuxième graphique les motifs de retard sans ce dernier.

Seconde visualisation

Analyse : On comprend grâce à ce second graphique que le trafic routier représente le premier motif de retard dans le cadre d'une intervention. Un motif également intéressant qui arrive en 3e position est les adresses incomplètes ou fausses, ce motif peut-être corrigé lors de l'appel de déclaration de l'incident.

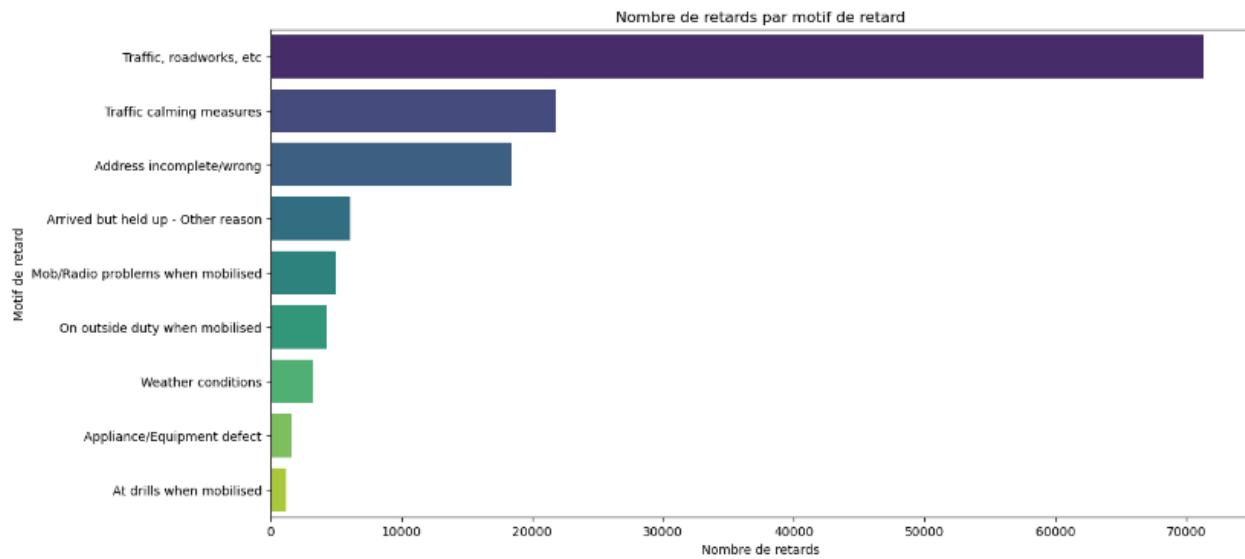


Figure 25 - Nombre de retards par motif de retard sans les "Not held up"

3. Analyses statistiques

Nous pouvons nous concentrer sur une analyse statistique de 2 variables du dataset : TravelTimeSeconds et TurnoutTimeSeconds, qui composent notre variable cible AttendanceTimeSeconds.

La variable TurnoutTimeSeconds a une moyenne 2,5 fois plus petite que la moyenne du TravelTimeSeconds. Cette variable est le temps de départ sur les lieux et doit évidemment être la plus rapide possible.

La séquence de temps TravelTimeSeconds est la séquence influençant le plus le temps d'intervention. Cette variable pourrait être améliorée en réfléchissant au choix des itinéraires mais compliqué à déterminer et à optimiser en cas d'interventions pendant des périodes de pointe du traffic routier.

Analyse statistique de la variable TravelTimeSeconds	
count	2.345259e+06
mean	2.740589e+02
std	1.513036e+02
min	0.000000e+00
25%	1.730000e+02
50%	2.470000e+02
75%	3.430000e+02
max	1.198000e+03
Name:	TravelTimeSeconds, dtype: float64

Figure 26 - Données de la variable TravelTimeSeconds

```
Analyse statistique de la variable TurnoutTimeSeconds
count      2.345347e+06
mean       7.963973e+01
std        4.453185e+01
min        0.000000e+00
25%        5.500000e+01
50%        7.400000e+01
75%        9.600000e+01
max        1.192000e+03
Name: TurnoutTimeSeconds, dtype: float64
```

Figure 27 - Données de la variable TurnoutTimeSeconds

VI) Nettoyage et Pre-processing

Suite aux constats établis lors des étapes précédentes, nous allons maintenant nettoyer notre jeu de données, et si besoin le transformer et l'enrichir.

L'objectif à la fin de cette étape est d'avoir un dataset prêt pour effectuer des analyses approfondies et une modélisation à l'aide de Machine Learning.

1. Apprentissage supervisé ou non supervisé?

Il s'agit d'un jeu de données avec modèle à apprentissage supervisé car les données sont labellisées.

On cherche à prédire une variable cible identifiée : le temps d'arrivée sur les lieux de l'incident.

2. Régression ou classification?

La variable cible , temps d'arrivée sur les lieux de l'incident, étant une variable quantitative, il s'agit d'un problème de régression.

Notre variable cible pourra être isolée dans un dataframe “target”.

Nom de la colonne	Description	Type	Taux de NA
AttendanceTimeSeconds	Temps d'arrivée sur les lieux en sec	int64	0%

3. Processus de Nettoyage et de traitement des données

a. Identification des doublons

Il n'y a aucun doublon dans le jeu de données “Incidents”.

Il y a effectivement des doublons dans le jeu de données “Mobilisations” sur la variable “IncidentNumber”. Cela s'explique car il peut y avoir plusieurs équipes mobilisées sur un même incident.

b. Identification des valeurs manquantes

Dans les 2 jeux de données “Incidents” et “Mobilisation”, Il y a quelques données avec des NaN, que nous allons pouvoir transformer de différentes manières :

- Ajout d'une nouvelle classe pour les éléments “nuls” mais qui sont en fait “normaux” car non concernés par la variable.
- Imputer avec la méthode SimpleImputer de stratégie “mediane” sur les valeurs numériques.

- Diviser les dates en année, mois, jour, jour de la semaine, heure
- Convertir certaines colonnes en d'autres types.
- Utiliser la méthode bfill sur les variables catégorielles ayant peu de valeurs nulles, le mode si la variable comporte beaucoup de valeur nulle.

On pourra distinguer plus précisément ces transformations dans le paragraphe 4.

c. Variables explicatives conservées

i. Jeu de données "Incidents"

Dans ce jeu de données, nous pouvons conserver les variables explicatives suivantes :

Nom de la colonne	Description
TimeOfCall	Heure d'appel de déclaration de l'incident
IncidentGroup	Catégorie principale de l'incident
StopCodeDescription	Catégorie détaillée de l'incident
SpecialServiceType	Type de service spécial
PropertyCategory	Catégorie de la propriété sur/dans laquelle à lieu l'incident (résidence, structure ext..)
Postcode_district	Code postal du quartier
IncGeo_BoroughName	Nom de l'arrondissement
IncidentStationGround	Nom de la caserne qui est intervenue
FirstPumpArriving_AttendanceTime	Temps d'arrivée de la première équipe de secours (en seconde)
FirstPumpArriving_DeployedFromStation	Caserne qui a déployé la première équipe de secours
SecondPumpArriving_AttendanceTime	Temps d'arrivée de la seconde équipe de secours (en seconde)
SecondPumpArriving_DeployedFromStation	Caserne qui a déployé la seconde équipe de secours
NumStationsWithPumpsAttending	Nombre de casernes avec secours déployés
NumPumpsAttending	Nombre d'équipes avec secours déployés réels
PumpCount	le nombre réel de pompes à incendie ou de moteurs déployés pour chaque incident.
PumpMinutesRounded	Arr. du nombre de min pendant laquelle les pompes à incendie ont été utilisées lors d'un incident.
NumCalls	Nombre d'appels

ii. Jeu de données “Mobilisations”

Dans ce jeu de données, nous pouvons conserver les variables explicatives suivantes :

Nom de la colonne	Description
HourofCall	Heure d'appel de déclaration de l'incident
Resource_Code	code des équipages mobilisés
PerformanceReporting	indicateur de performance de prise en charge de l'incident
DateAndTimeMobilised	Heure et date du début de la prise en charge de l'incident
DateAndTimeMobile	Date et heure de départ pour les lieux de l'incident
DateAndTimeArrived	heure d'arrivée sur les lieux de l'incident
TurnoutTimeSeconds	temps de départ pour les lieux de l'incident en sec
TravelTimeSeconds	temps de trajet en sec
DateAndTimeLeft	date et heure de départ du lieu d'incident
DeployedFromStation_Name	nom de la caserne
DeployedFromLocation	localisation géographique de départ de l'unité déployée
PumpOrder	Nombre de camion demandé sur les lieux
PlusCode_Code	ID du Type de prise en charge
DelayCode_Description	Description du motif de retard

d. Séparation des types de variables

Nous avons choisi de séparer nos variables en 2 types : numériques et catégorielles.

Variables numériques :

```
num_feats = ["HourOfCall", "FirstPumpArriving_AttendanceTime",
"NumStationsWithPumpsAttending", "NumPumpsAttending", "PumpCount",
"PumpMinutesRounded", "NumCalls", "TurnoutTimeSeconds", "TravelTimeSeconds",
"SecondPumpArriving_AttendanceTime", "PumpOrder"]
```

Variables catégorielles :

```
cat_feats = ["TimeOfCall", "IncidentGroup", "StopCodeDescription", "PropertyCategory",
".PropertyType", "IncGeo_BoroughName", "FirstPumpArriving_DeployedFromStation",
"SecondPumpArriving_DeployedFromStation",
"PerformanceReporting", "DeployedFromStation_Name", "PlusCode_Code",
"SpecialServiceType", "Postcode_district", "IncidentStationGround",
"DelayCode_Description", "DeployedFromLocation"]
```

4. Transformation des données

a. Jeu de données “Incidents”

Nom de la colonne	Description	Type informatique	Taux de NA 2009 to 2017	Taux de NA 2018 onwards	Distribution des valeurs	Gestion des NaN	Encodage?
TimeOfCall	Heure d'appel de déclaration de l'incident	object	0,00%	0,00%	Catégorielle - sup. à 10 catégories	non	dtype 'object' à changer à 'datetime' puis à séparer en jour, mois, année, heure, min, sec
IncidentGroup	Catégorie principale de l'incident	object	0,00%	0,00%	Catégorielle - 3 à 5 catégories	non	Encodage CAT : get.dummies ou OneHotEncoder
StopCodeDescription	Catégorie détaillée de l'incident	object	0,00%	0,00%	Catégorielle - 5 à 10 catégories	non	Encodage CAT : get.dummies ou OneHotEncoder
SpecialServiceType	Type de service spécial	object	69,74%	65,00%	Catégorielle - sup. à 10 catégories	oui transformer les Nan en "no special service"	Encodage CAT
PropertyCategory	Catégorie de la propriété sur/dans laquelle à lieu l'incident (résidence, structure ext.)	object	0,00%	0,00%	Catégorielle - 5 à 10 catégories	non	Encodage CAT
Postcode_district	Code postal du quartier	object	0,00%	0,00%	Catégorielle - sup. à 10 catégories	non	Encodage CAT
IncGeo_BoroughName	Nom de l'arrondissement	object	0,00%	0,00%	Catégorielle - sup. à 10 catégories	non	Encodage CAT
IncidentStationGround	Nom de la caserne qui est intervenue	object	0,00% 0,000140 %		Catégorielle - sup. à 10 catégories	non	Transformation en liste de nombres car 103 valeurs uniques?
FirstPumpArriving_AttendanceTime	Temps d'arrivée de la première équipe de secours (en seconde)	float64	9,25% 5.705050 %		Quantitative	oui moyenne sur les Nan avec SimpleImputer	StandardScaler
FirstPumpArriving_DeployedFromStation	Caserne qui a déployé la première équipe de secours	object	9,25% 5.705891 %		Catégorielle - sup. à 10 catégories		Encodage CAT
SecondPumpArriving_AttendanceTime	Temps d'arrivée de la seconde équipe de secours (en seconde)	float64	64,95% 63.243936 %		Quantitative	oui, si pas de seconde équipe alors 0	StandardScaler
SecondPumpArriving_DeployedFromStation	Caserne qui a déployé la seconde équipe de secours	object	64,95% 63.244637 %		Catégorielle - sup. à 10 catégories	oui, si pas de seconde équipe alors "no second pump required"	Encodage CAT
NumStationsWithPumpsAttending	Nombre de casernes avec secours déployés	float64	0,56%	1.073920 %	Quantitative	oui moyenne sur les Nan avec SimpleImputer	StandardScaler
NumPumpsAttending	Nombre d'équipes avec secours déployés réels	float64	0,56%	1.073920 %	Quantitative	oui moyenne sur les Nan avec SimpleImputer	StandardScaler
PumpCount	le nombre réel de pompes à incendie ou de moteurs déployés pour chaque incident.	int64	0,00%	0,00%	Quantitative	non	StandardScaler
PumpMinutesRounded	Arr. du nombre de min pendant laquelle les pompes à incendie ont été utilisées lors d'un incident.	int64	0,00%	0,00%	Quantitative	non	StandardScaler
NumCalls	Nombre d'appels	float64	0,18% 0,002523 %		Quantitative	oui moyenne sur les Nan avec SimpleImputer	StandardScaler

Figure 28 - Tableau explicatif de la transformation des données du jeu “Incidents”

b. Jeu de données “Mobilisations”

Nom de la colonne	Description	Type informatique	Taux de NA mobi_2009	Taux de NA mobi_2015	Taux de NA mobi_2021	Distribution des valeurs	Gestion des NaN	Encodage / Transformation
HourOfCall	Heure d'appel de déclaration de l'incident	int64 convert en date time	0,00%	0,00%	0,00%		non	non
Resource_Code	code des équipages mobilisés	object	0,00%	0,00%	0,00%		non	Encodage CAT
PerformanceReporting	indicateur de performance de prise en charge de l'incident	object	0,00%	0,00%	0,00%	Catégorielle - Binaire 1, 2	non	non
DateAndTimeMobilised	Heure et date du début de la prise en charge de l'incident	datetime64[ns]	0,00%	0,00%	0,00%	Quantitative	non	Transformation : 1 colonne par jour/mois/année/heure/min/sec
DateAndTimeMobile	Date et heure de départ pour les lieux de l'incident	datetime64[ns]	2,04%	0,82%	0,43%	Quantitative	oui SimpleImputer NUM ou bfill méthode	Transformation : 1 colonne par jour/mois/année/heure/min/sec
DateAndTimeArrived	heure d'arrivée sur les lieux de l'incident	datetime64[ns]	0,00%	0,00%	0,00%	Quantitative	non	Transformation : 1 colonne par jour/mois/année/heure/min/sec
TurnoutTimeSeconds	temps de départ pour les lieux de l'incident en sec	float64	2,05%	0,83%	0,43%	Quantitative	oui SimpleImputer NUM	non
TravelTimeSeconds	temps de trajet en sec	float64	2,05%	0,84%	0,43%	Quantitative	oui SimpleImputer NUM	non
DateAndTimeLeft	date et heure de départ du lieu d'incident	datetime64[ns]	5,11%	0,11%	0,04%	Quantitative	oui SimpleImputer NUM	StandardScaler NUM
DeployedFromStation_Name	nom de la caserne	object	0,00%	0,02%	0,00%	Catégorielle - sup. à 10 catégories	non	Encodage CAT
DeployedFromLocation	localisation géographique de départ de l'unité déployée	object	0,00%	0,08%	0,08%	Catégorielle - Binaire	non	Encodage CAT
PumpOrder	Nombre de camion demandé sur les lieux	int64	0,00%	0	0,00%	Catégorielle - sup. à 10 catégories	non	StandardScaler NUM
PlusCode_Code	ID du Type de prise en charge	object	0,00%	0,00%	0,00%	valeur unique initial	non	Encodage CAT
DelayCode_Description	Description du motif de retard	object	76,22%	75,58%	75,27%	Catégorielle - 5 à 10 catégories, 'Not held up', 'Address incomplete/wrong', 'Traffic, roadworks, etc', 'Traffic calming measures', 'Arrived but held up - Other reason', 'Appliance/Equipment defect', 'Mobi/Radio problems when mobilised', 'On outside duty when mobilised', 'Weather conditions', 'At drills when mobilised'	oui > transformer les "0" en "on time"	Encodage CAT

Figure 29 - Tableau explicatif de la transformation des données du jeu “Mobilisations”

VII) Conclusion et projection sur la partie modélisation.

Notre étude dans cette première partie a porté grandement sur l'analyse des données et leurs pertinences. Cela dans l'optique de préparer à la réalisation d'un modèle prédisant au mieux le temps d'arrivée sur les lieux des pompiers aussi appelé temps d'intervention.

Ce d'autant plus que les variables disponibles dans les 5 datasets étaient nombreuses et certaines fois difficiles à bien en saisir la description. Toutefois, nous pouvons signaler qu'il était très intéressant d'avoir autant de données précises et de variables descriptives recensant les incidents et les interventions des pompiers.

L'analyse des données nous a donc permis de bien comprendre et de nous familiariser avec les variables.

Par la suite, nous avons ciblé des graphiques à réaliser en fonction de grandes catégories définissant nos variables et pouvant avoir un impact sur la variable cible ou apporter du contexte.

La DataViz nous a donc permis d'identifier des tendances, les facteurs qui influencent le temps d'intervention ou d'écartez au contraire certains qui n'apportent que peu de variation sur la variable cible.

Finalement, nous avons défini les variables explicatives à conserver afin de prédire la variable cible "AttendanceTimeSeconds" dans le cadre d'une régression d'un modèle d'apprentissage supervisé.

Variables que nous avons séparées entre les variables catégorielles et numériques. La question de la gestion du nettoyage des données s'est alors posée pour celles-ci. Nous avons donc défini les modalités de gestion des valeurs manquantes, de la transformation et de l'encodage des données.

Sur cette base, nous nous projetons vers la prochaine étape qui consistera à tester et développer différents modèles prédictifs pour estimer au mieux notre variable cible.

Nous serons donc amenés à explorer différents modèles, les optimiser et identifier le plus performant.

ETAPE 2 - Modélisation

I) Nos étapes d'itérations de modélisation

Pour parvenir à un résultat de modélisation de régression que nous jugeons satisfaisant (métrique R^2 proche de 1 et erreur MAE proche de 15), nous avons commencé tout d'abord par une approche classique, avant de réfléchir à une optimisation du modèle par feature engineering.

Voici les étapes par lesquelles nous sommes passées :

- 1) Recherche des corrélations entre variables
- 2) Choix des colonnes à conserver
- 3) Gestion des doublons
- 4) Gestions des NaN par : transformation de données par la moyenne ou par le mode ou ajout de texte pour les données non renseignées
- 5) Réduction du nombre de lignes du dataframe pour améliorer la rapidité d'exécution du code grâce à la méthode sample de pandas.
- 6) Séparation des données labellisées et de la variable cible
- 7) Séparation des données numériques, temporelles, catégorielles et géographiques
- 8) Création de Pipelines par type de données
- 9) Recherche des hyper-paramètres des modèles
- 10) Tests de différents modèles avec différentes variables explicatives intégrées dans le dataframe :
 - a) Régession de Lasso
 - b) Régession linéaire
 - c) Régession de Ridge
 - d) ElasticNet
 - e) Random Forest
 - f) Gradient Boost Regressor
 - g) Decision Tree Regressor
 - h) SVR model
- 11) Comparaison de leurs métriques : R^2 , RMSE, MAE, MSE et MedAE, des résidus et des erreurs de prédictions

Malheureusement, nous nous sommes rendus compte qu'avec cette première méthode, les métriques des divers modèles n'étaient pas concluants. car les R^2 étaient tous trop proches de 1. Nous avons suspecté une fuite de données entraînant un surapprentissage.

Il fallait alors approfondir le pré-processing et le feature engineering pour mieux comprendre pourquoi nos modélisations n'étaient pas convenables. Nous avons donc engagé une nouvelle démarche de pré-processing et de feature engineering.

Nous avons commencé par créer un nouveau fichier comprenant les coordonnées géographiques, la catégorie géographique (urbain, banlieue, centre ville) et la zone géographique (nord, sud, est, ouest...). Nous avons pu ainsi créer une carte des stations de pompier et les incidents de notre jeu de données.

L'objectif étant de voir si la segmentation et les caractéristiques géographiques des stations pouvaient influer sur le temps de réponse. Nous n'avons conservé que les individus dont les coordonnées géographiques étaient renseignées.

Nous avons créé de nouvelles variables comme la relation entre la distance entre l'incident et la station, le temps moyen de réponse par station ou le nombre d'incident par zone géographique.

Nous avons ensuite traité les variables temporelles en segmentant les moments de la journée durant lesquelles avaient lieu les incidents (Night, Morning, early morning...), en indiquant les jours de weekend et en segmentant par saison. Nous avons procédé à un encodage cyclique de ces variables pour une meilleure cohérence.

Puis nous avons abordé les variables relatives à la criticité des incidents et aux retards, en créant un indice de criticité sur 20 pour chaque incident. Nous avons également intégré une variable binaire en cas de retard ou en cas d'arrivée d'un second camion ainsi que la variable d'interaction entre la distance entre la station et le lieu d'incident et la moyenne des temps de réponse en fonction du moment de la journée.

Nous avons également revu notre gestion des outliers en les supprimant pour les variables "AttendanceTimeSeconds", "TurnOutTimeSeconds", "TravelTimeSeconds".

II) Interprétation et optimisation des modèles de Machine Learning

Parmi les étapes citées précédemment, nous avons pu comparer différents modèles de régression et leurs métriques, afin de choisir la meilleure modélisation..

Nous nous sommes donnés pour objectif les métriques suivants :

- **R² > entre 0,7 et 0,9**
- **MAE > entre 30 et 60.**

1) Première approche : interprétation

Cette approche était une approche standard de pré-processing.

Nous avons fait varier :

- le nombre de variables explicatives,
- la façon de gérer les NaN des données numériques par la médiane ou la moyenne.

Voici les comparaisons des métriques des différents modèles testés avec nos premières hypothèses.

A cause du manque de puissance de calcul et d'un temps de réponse particulièrement long, nous avons pris le parti de sélectionner 200 000 lignes du data frame choisies aléatoirement grâce à la méthode sample de pandas.

a) Avec les variables "TurnoutTimeSeconds", "TravelTimeSeconds", "NumCalls", "PumpMinutesRounded"

Nom du modèle	Nan des num	alpha	RMSE	R ²	MAE	MSE	MedAE
Lasso	median	0,01	16,479	0,985	1,442	271,566	0,247
		0,1	15,996	0,986	1,864	255,858	0,783
		0,2	16,006	0,986	1,802	256,206	0,707
		0,09	15,994	0,986	1,872	255,823	0,785
		0,08	15,994	0,986	1,880	255,796	0,793
		0,07	15,993	0,986	1,891	255,792	0,801
		0,06	15,993	0,986	1,906	255,761	0,806
		0,05	15,988	0,986	1,921	255,621	0,797
		2	16,381	0,986	3,065	268,339	1,562
Ridge	median	0,01	16,870	0,985	2,451	284,590	0,600
		0,1	16,837	0,985	2,430	283,499	0,593
		0,001	16,873	0,985	2,458	284,698	0,606
	mean	0,01	18,028	0,983	2,618	324,996	0,629
		0,1	17,989	0,983	2,601	323,606	0,638
		0,2	17,943	0,983	2,612	321,954	0,680
		0,5	17,819	0,983	2,685	317,501	0,808
		0,7	17,749	0,983	2,732	315,016	0,881
		1,2	17,610	0,983	2,818	310,125	0,990
		2,5	17,388	0,984	2,925	302,348	1,150

Lasso		0,1	12,937	0,991	1,712	167,374	0,881
Random Forest		0,1	9,539	0,995	1,136	90,998	0,300

- b) SANS les variables hyper-corrélées à la variable cible :
 "TurnoutTimeSeconds", "TravelTimeSeconds"

Nom du modèle	Nan des num	alpha	RMSE	R ²	MAE	MSE	MedAE
Random Forest	mean	0,1	136,974	0,006	96,974	18762,007	70,317
Random Forest		0,1	90,439	0,562	44,791	8179,295	11,421
Lasso		0,1	88,485	0,580	46,447	7829,630	14,566
Gardient Boost Regressor		0,1	89,228	0,573	45,014	7961,629	13,662
Ridge		0,1	88,455	0,581	46,416	7824,266	14,502
DecisionTreeRegressor	median	-	101,192	0,458	48,410	10239,771	9,000

- c) SANS les variables hyper-corrélées à la variable cible :
 "FirstPumpArriving_AT" et "SecondPumpArriving_AT"

Nom du modèle	Nan des num	alpha	RMSE	R ²	MAE	MSE	MedAE
DecisionTreeRegressor	Median	-	148,961	-0,175	99,706	22189,528	66,0

2) Seconde approche : optimisation

a) Feature engineering et pré-processing

Après avoir repris le fichier du début et refait du feature engineering sur les données géographiques, les données temporelles des incidents, les retards et la gravité des incidents et sur les pump attendance time (first et seconds), nous sommes arrivés à la heatmap suivante.

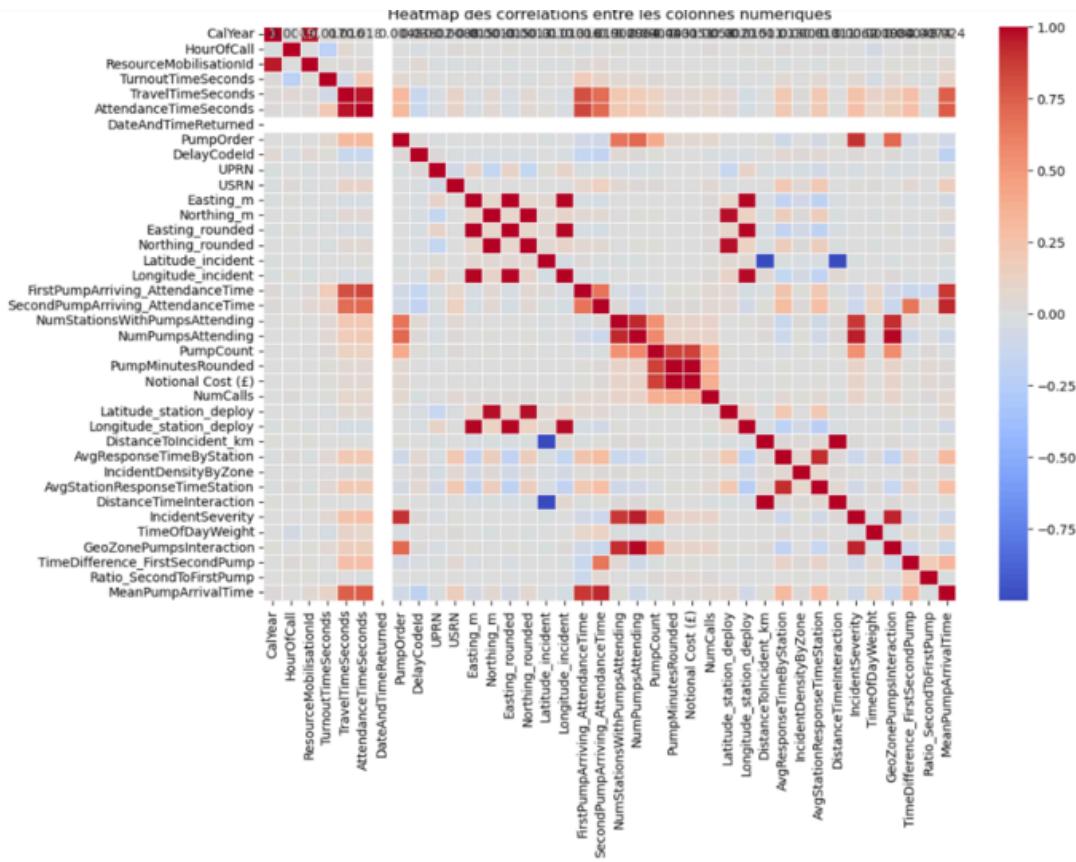


Figure 30 - Heatmap des corrélations entre les colonnes numériques

Nous observons donc un peu plus de features corrélées avec notre variable cible AttendanceTimeSeconds et plus de capture des corrélations entre les variables explicatives.

Ensuite, nous avons cherché les caractéristiques les plus importantes pour la variable cible :

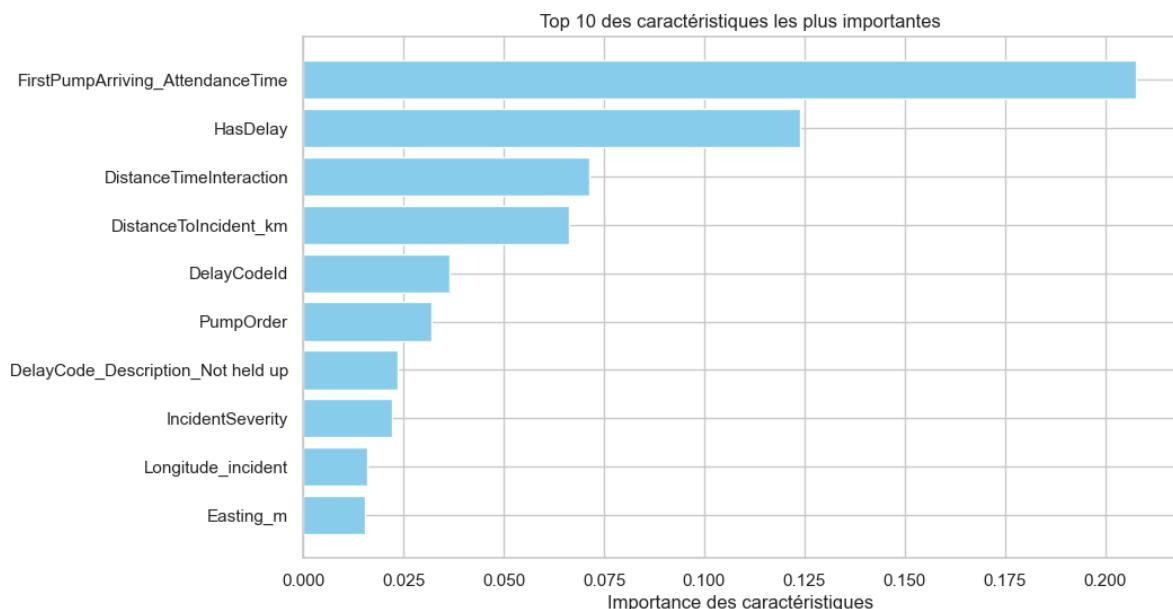


Figure 31 - Graphique en barres des 10 features les plus importantes

b) Gestion des outliers corrélées à la variable cible

Puis, pour voir si une gestion des outliers via une suppression des valeurs temporelles (variable cible “AttendanceTimeSeconds” et des variables composantes de celle-ci “TurnOutTimeSeconds” et “TravelTimeSeconds”) au lieu d'une attribution des bornes inférieures ou supérieures aurait un impact sur les résultats du modèle, nous avons supprimé ces outliers dans le dataframe initial juste avant la transformation en un dataframe de modélisation.

Cela a permis notamment d'éliminer des lignes incohérentes (ie. des cas où le temps de préparation était supérieur au temps d'intervention).

Aussi, le nombre de lignes supprimées par cette gestion était faible sur le volume total et nous permettait de partir dans cette démarche.

Nous avons pu observer de nettes améliorations de performances du modèle après ce pré-processing. C'est pourquoi nous avons conservé cette transformation.

Voici les métriques des différents modèles après gestion des outliers “AttendanceTimeSeconds”, “TurnOutTimeSeconds” et “TravelTimeSeconds” :

Nom du modèle	alpha	R ²		RMSE	MAE	MSE	MedAE
Random Forest	-	on passe de 0.8186 => 0.8149	0,810	50,784	34,048	2578,989	22,125
Lasso	0,33	on passe de 0.8657 => 0.8787	0,879	40,623	24,941	1650,261	15,877
	0,695	on passe de 0.8640 => 0.8756	0,876	41,102	24,543	1689,355	15,216
ElasticNet	-	on passe de 0.6814 => 0.8780	0,878	40,738	26,103	1659,550	17,470
SVR / SVM model	-	on passe de 0.6058 => 0.8483	0,848	45,436	28,724	2064,408	18,259

La variable FirstPumpArriving_AttendanceTime étant l'une des features les plus importantes, nous avons testé également ce retrait d'outliers sur celle-ci. Cependant, cela dégradait le modèle. Nous n'avons donc pas conservé cette transformation.

c) Approche avec l'ACP

Ensuite, nous avons testé une approche avec l'ACP mais cela n'accélère pas le temps de calcul du modèle et réduit les performances de nos métriques. Nous avons donc écarté cette solution..

d) Encodage des variables cycliques

Finalement, nous avons essayé un encodage cyclique sur les variables : "DayOfWeek", "Month", "HourOfCall", "TimeOfDay".

Cela n'a pas amélioré de manière significative les performances de notre modèle. Ce qui indique que ces variables n'ont pas de relation forte avec la variable cible.

Après avoir complété la heatmap avec les caractéristiques cycliques (HourOfCall_sin, HourOfCall_cos, DayOfWeek_sin, DayOfWeek_cos, Month_sin et Month_cos), la corrélation avec "AttendanceTimeSeconds" semble être très faible. Cela aussi suggère que ces caractéristiques n'influencent peut-être pas fortement notre variable cible.

Voici les métriques des différents modèles après encodage des variables cycliques :

Nom du modèle	alpha	R ²		RMSE	MAE	MSE	MedAE
Random Forest		on passe de 0.8149 => 0.8131	0,813	47,420	32,437	2248,645	21,596
Lasso	0,330	on passe de 0.8787 => 0.8597	0,860	41,081	24,973	1687,617	15,724
	0,695	on passe de 0.8756 => 0.8552	0,855	41,743	24,619	1742,472	14,962
ElasticNet		on passe de 0.8780 => 0.8602	0,860	41,015	25,970	1682,223	17,341
SVR / SVM model		on passe de 0.8483 => 0.8268	0,827	45,617	27,241	2080,912	16,369

Nous n'observons pas d'améliorations, voire une dégradation du modèle après gestion des outliers "AttendanceTimeSeconds", "TurnOutTimeSeconds" et "TravelTimeSeconds" .

Finalement, dans notre sélection finale, choisissant aléatoirement 200 000 lignes du data frame parmi les années, nous avons conservé cet encodage

hormis sur la variable “TimeOfDay” afin d’éviter la redondance d’informations dans le modèle.

Bien que les caractéristiques cycliques ne semblent pas montrer une forte corrélation directe avec la variable cible, elles pourraient néanmoins offrir des avantages en aidant à capturer des schémas et des interactions dépendants du temps.

e) Sélection aléatoire des lignes

Finalement, afin de pouvoir analyser l’impact des données cycliques sur le modèle nous avons sélectionné de manière aléatoire 200 000 lignes du dataframe initial qui contenait toutes les années.

3) Comparaisons des métriques des modèles testés

Voici les comparaisons des métriques des différents modèles testés après la fin des optimisations :

Nom du modèle	alpha	R ²	RMSE	MAE	MSE	MedAE
Random Forest	-	0,810	50,784	34,048	2578,989	22,125
Lasso	0,33	0,879	40,623	24,941	1650,261	15,877
	0,695	0,876	41,102	24,543	1689,355	15,216
ElasticNet	-	0,878	40,738	26,103	1659,550	17,470
SVR / SVM model	-	0,848	45,436	28,724	2064,408	18,259

Les 2 meilleurs modèles avec des métriques très proches sont :

- Lasso avec paramètre alpha de 0.33
- ElasticNet

4) Analyse graphique des modèles testés

Pour valider le choix final de l’un de ces modèles nous avons ensuite réalisé les graphes suivants afin d’analyser les erreurs de prédictions.

- une comparaison des valeurs réelles vs prédictes (Elastic Net)

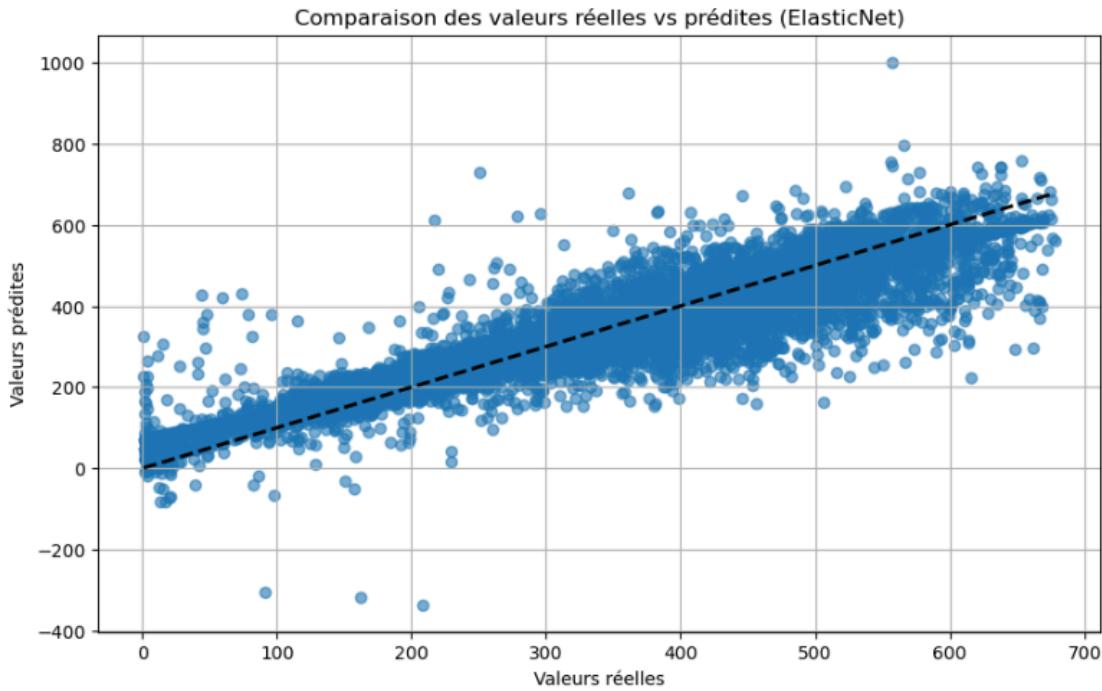


Figure 32 - Graphique de comparaison des valeurs réelles vs prédictes (Elastic Net)

- la fréquence de la distribution des erreurs (résidus) (Elastic Net)

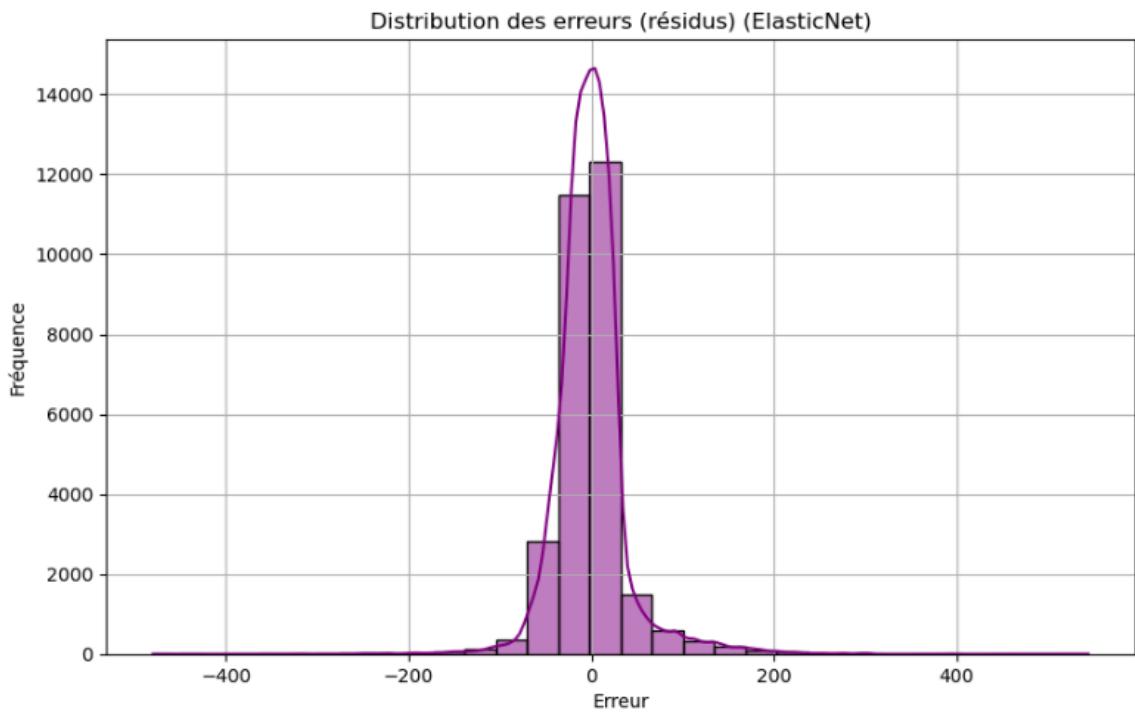


Figure 33 - Graphique de la distribution des erreurs (résidus) du modèle Elastic Net

- l'analyse des résidus des valeurs prédictes (Elastic Net)

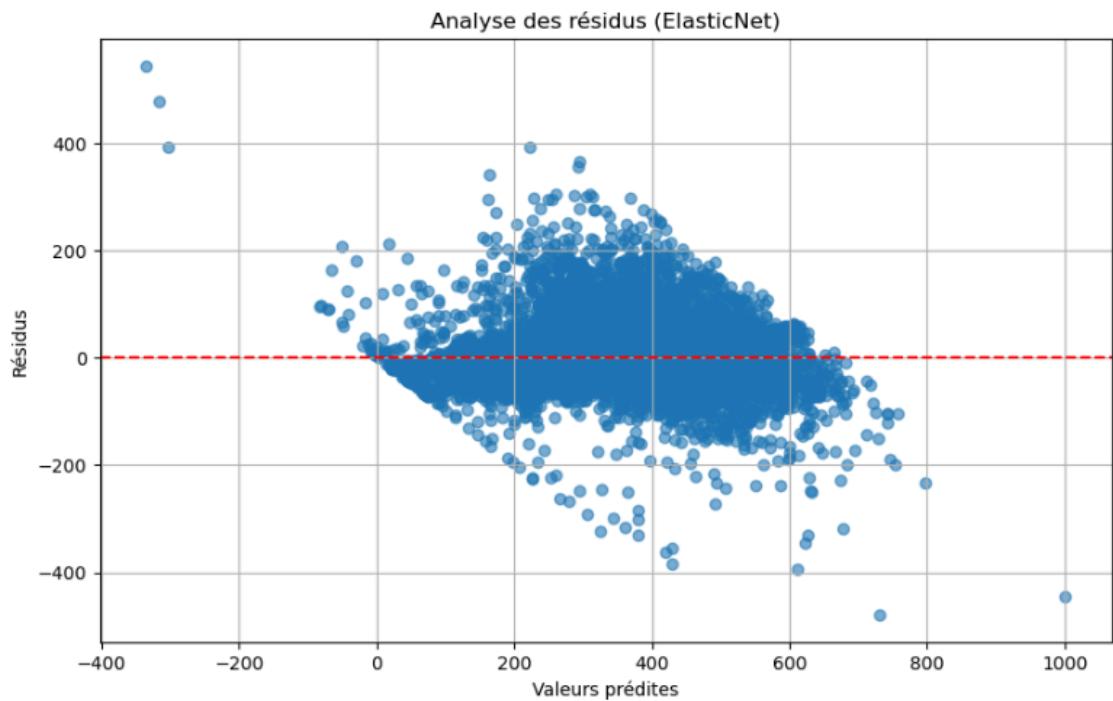


Figure 34 - Graphique d'analyse des résidus du modèle Elastic Net

- une courbe des résidus cumulés (Elastic Net)

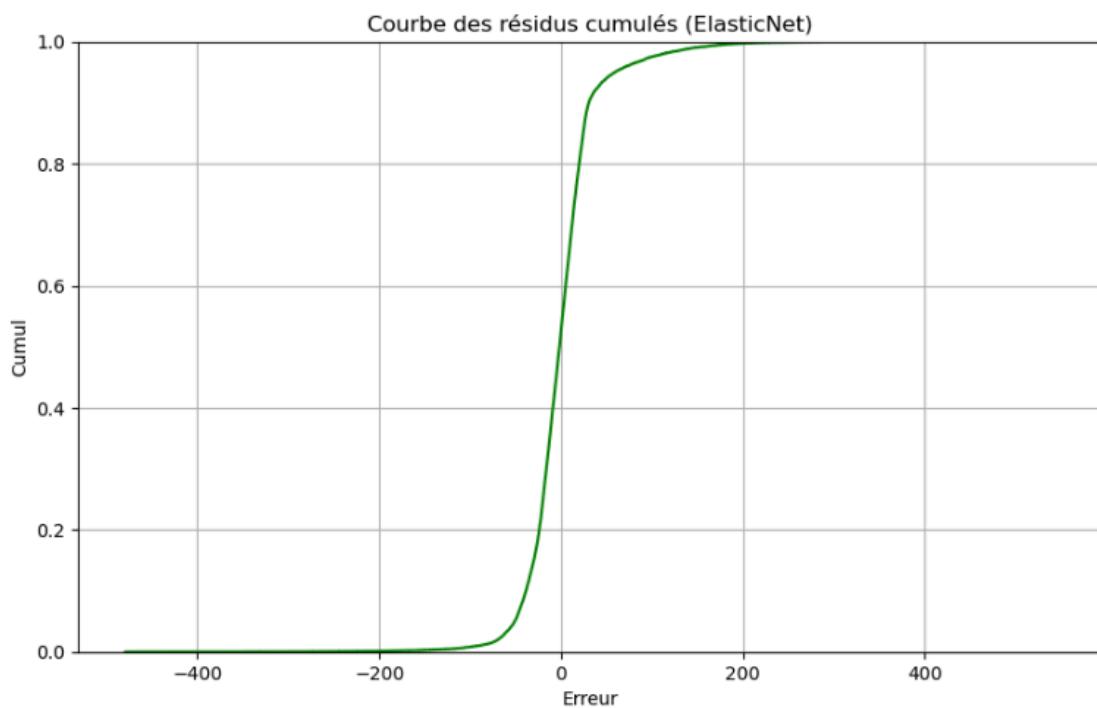


Figure 35 - Graphique d'analyse des résidus cumulés du modèle Elastic Net

IV) Conclusions scientifiques et métiers de la modélisation

Conclusion scientifique

Nous pouvons voir que le modèle amenant les meilleures performances est **Elastic Net avec pour hyperparamètres l1 = 0.9 et alpha =0.233**, avec les métriques d'évaluation suivantes :

ElasticNet Regression Performances :

RMSE: 40.30510232051049

R²: 0.8872394816483825

MAE: 26.074362160775582

MSE: 1624.5012730668202

MedAE: 18.116531109688907

- Les prédictions du modèle s'écartent en moyenne de 40.30 unités (secondes) des valeurs réelles selon la RMSE.
- 88,7% de la variance des données est expliquée par le modèle.
- La MAE de 26.07 signifie que les prédictions du modèle s'écartent en moyenne de 26.07 unités (secondes).
- La moitié des erreurs absolues sont inférieures à 18.116 unités.

On peut en conclure que le modèle est bien ajusté et fait des prédictions assez précises.

Avant la mise en production de ce modèle, il serait utile de lui soumettre de nouvelles données afin d'être certains que le modèle reste performant.

Conclusion métier

Le modèle ElasticNet ($\text{l1} = 0.9$; $\text{alpha} = 0.33$) choisi pourrait permettre à la brigade des pompiers de Londres de prévoir leur temps de réponse aux incidents par heure, jour, mois et selon la zone géographique, à 26 secondes près.

Nous considérons donc ce modèle utilisable et utile sur de nouvelles données notamment sur celles actualisées et à venir de 2024. C'est donc intéressant d'un point de vue métier de vérifier les performances du modèle sur de nouvelles données et constater une amélioration ou une baisse de performance opérationnelle. Voici les principaux avantages:

Allocation Optimisée des Ressources

La recherche et l'optimisation du modèle qui a été nécessaire lors de nombreuses étapes nous a permis de mieux analyser les facteurs qui influencent les résultats.

Ces prédictions pourront permettre à la LFB d'améliorer leurs ressources et leurs organisations.

Gestion des Incidents

Disposer de prévisions proches fait donc le lien avec les KPI de la London Fire Brigade pour atteindre ses objectifs économiques et de service, et pourra aider à la décision.

Que cela soit pour optimiser le délai d'intervention moyen, le temps d'arrivée de la première et seconde équipe ou encore sur la réduction des dégâts matériels et pertes humaines.

Amélioration de la Sécurité Publique

En s'appuyant sur ce modèle Elastic Net, la LFB peut améliorer sa planification opérationnelle, réduire les temps de réponse et améliorer la sécurité publique.

Cette approche basée sur les données permettra à la brigade d'anticiper les défis et de les résoudre avant qu'ils ne deviennent critiques, garantissant ainsi des réponses d'urgence plus efficaces et rapides à travers Londres.

ETAPE 3 - Conclusions projet

I) Difficultés rencontrées lors du projet

Pendant la réalisation du projet, le principal verrou scientifique rencontré a été d'avoir réalisé un nettoyage et un pré-processing qui nous semblait complet et qui nous menait à des métriques non réalistes (notamment R^2 très proche de 1, de l'ordre de 0.98, 0.99). Et de ne plus savoir comment améliorer ce modèle.

Les tâches qui nous ont prises plus de temps que prévu ont été :

- Prévisionnel : Nous avons toujours réussi à atteindre les jalons projet fixés. La data visualisation et la modélisation, comme expliqué ci-dessous nous ont occupées plusieurs semaines, qui auraient pu nous pénaliser.
- Jeux de données : Au début du projet pendant quelques semaines, nous nous étions trompés sur la variable cible et en avions créé une nouvelle, alors que cette variable cible existait déjà dans les colonnes initiales. Cela nous a fait perdre du temps sur la Data visualisation que nous avons dû relancer et parfois revoir complètement en utilisant la bonne variable.
- Compétences techniques / théoriques : Il nous fallu de l'aide et l'expérience du chef de projet DataScientest pour comprendre qu'il fallait que l'on gère également les outliers de la variable cible et des variables explicatives hyper corrélées, afin d'avoir un modèle final satisfaisant.
- Pertinence : Une question s'est posée concernant l'utilisation du modèle que nous avions mis en place, qui impliquait une sélection des variables explicatives. Premier cas de figure, le modèle était utilisé "sur le moment" par la brigade des pompiers de Londres et en fonction des coordonnées GPS de l'incident et de la station pour prédire en temps réel le temps d'intervention. Dans ce cas précis, nous ne pouvions pas utiliser les variables temporelles liées à l'incident en cours. Dans un second cas de figure, notre modèle était utilisé à des fins de fixation des objectifs par exemple à savoir si les temps d'intervention effectués sur l'année correspondent aux prévisions du modèle. Nous avons choisi la seconde option, car les variables explicatives liées à l'incident en cours comme "First Pump Attendance Time Seconds" améliorent considérablement les prévisions de notre modèle sans peser trop lourd en termes de feature importance.
- IT : nous avons dû réduire le jeu de données à 1 année, pour réaliser nos modélisations car cela prenait trop de temps de chargement avec l'ensemble des données. De plus, nous n'avons pas réussi à utiliser un outil de code collaboratif tel que Google Collab. Nous nous sommes partagés par Github ou WeTransfer un Jupyter Notebook que nous complétions chacun au fur et à mesure. Ceci aurait pu amener à des doublons ou des erreurs et blocages. Les points réguliers que nous avons réalisés en équipe, nous ont permis d'éviter cela.

II) Bilan

1. Contribution principale dans l'atteinte des objectifs projet

Paul : Ma principale contribution a été le nettoyage, le pré-processing et le feature engineering lors de notre deuxième approche, ainsi que la construction des modèles et des itérations de test. J'ai également contribué à la partie data visualisation et exploration du jeu de données et encodage.

Marie : Ma principale contribution a été la gestion organisationnelle des réunions de groupe et le cadencement des actions pour tenir les délais, ainsi que la préparation de la présentation Streamlit. Bien évidemment, afin d'apprendre, j'ai également travaillé sur l'exploration des données, la data visualisation, le pré-processing et le machine learning.

Thomas : J'ai participé principalement au processus d'exploration des données et à la data visualisation qui s'en est suivi. J'ai également contribué lors de la phase de pré-processing et en testant les modèles précédemment créés par l'équipe en appliquant une gestion des outliers.

Namita: J'ai participé principalement en compréhension et manipulation de données, en identification des relations entre variables, explicatives et cibles, quelques graphiques en data visualisation. Tester les modèles créés par mes collègues avec une autre approche : modification du pre-processing pour vérifier si ça change le résultat.

2. Résultats obtenus comparés au benchmark initial

Initialement, nous avons vu que la brigade des pompiers de Londres avait un temps moyen de réponse appelé également temps d'arrivée sur les lieux d'un incident ("AttendanceTimeSeconds") d'environ **350 secondes**.

Grâce à notre modèle nous constatons que même pour des valeurs réelles inférieures ou supérieures à ce temps moyen de réponse les prédictions sont relativement fiables.

Finalement, la comparaison graphique entre les valeurs réelles et les valeurs prédites par le modèle retenu nous conforte dans la cohérence de ces résultats. En effet, la majorité des valeurs prédites sont concentrées en rapport des valeurs réelles et nous constatons assez peu de dispersion.

3. Atteinte des objectifs

Notre objectif projet était de prédire le temps de réponse à des interventions de la brigade des pompiers de Londres, fonction du nombre d'équipes déployées, du type d'incidents et du lieu de l'incident.

Nous sommes effectivement capables de prédire avec un intervalle de certitude de 26 secondes le temps d'intervention de la brigade des pompiers de Londres. Par la sélection de nos variables explicatives ainsi que les features importance de ces variables mais aussi grâce aux visualisations que nous avons pu mettre en place, nous sommes en mesure d'établir les facteurs qui influencent le plus négativement ce temps d'intervention:

- Naturellement les caractéristiques de la zone géographique où se situent la caserne et le lieu de l'incident créent de la variance dans le temps d'intervention, si la zone est très fréquentée : urbaine, les temps ont tendance à s'allonger du fait de la circulation.
- D'un point de vue général, plus on arrive en périphérie de la ville de Londres plus les temps d'intervention ont tendance à s'allonger.
- On observe également une forte concentration des incidents en hyper centre, malgré cela seulement 3 casernes peuvent trianguler efficacement les incidents qui ont lieu dans ce périmètre.
- On constate également que le nombre de fausses alarmes est très important notamment sur les incidents à forte criticité qui entraîne un nombre important de pompes déployées qui ne sont pas mobilisés sur des incidents à proximité.
- L'heure à laquelle est mobilisée la brigade joue également sur ce temps d'intervention puisqu'on constate une augmentation du temps de trajet entre 10h et 18h.
- On note dans les motifs de retard trois motifs reviennent, le trafic routier, et les travaux d'urbanisation, mais surtout les adresses fausses ou incomplètes, facteur sur lesquels les équipes pourront agir.

Nos métriques de modélisation pourront aider au process métier de réponse aux appels de la brigade des pompiers de Londres.

En effet, lors d'un incident, si une ou plusieurs équipes doivent intervenir, la plateforme d'appel pourra prévoir à xx secondes le temps d'arrivée sur les lieux. Ceci permettra de rassurer les citoyens et d'organiser plus rapidement les interventions entre les différentes équipes.

III) Suite du projet

Dans notre modélisation, pour des questions de temps et de performance, nous avons réduit notre jeu de données à 200 000 lignes contre 2,4 millions de lignes initialement. Pour augmenter les performances de notre modèle et l'affiner, nous aurions pu utiliser plus de données.

D'autre part, notre modèle se base sur des données de 2021 à avril 2024. Afin de tester le modèle, nous aurions pu lancer la prédiction avec les données actualisées de l'année 2024.

Il pourrait être donc intéressant pour la suite de vérifier et tester notre modèle avec ces nouvelles données sachant qu'elles sont actualisées tous les mois et mises à disposition sur le site Internet de la LFB.

Finalement, pour vérifier que nos prédictions étaient cohérentes avec les KPI imposés à la Brigade des pompiers de Londres, nous aurions pu redéfinir 4 modèles associés à des variables explicatives propres.

Ceci pour s'assurer de l'atteinte des indicateurs suivants :

- Temps moyen d'arrivée du premier équipage (mensuel) : **6 min**
- Temps moyen d'arrivée du second équipage (mensuel) : **8 min**
- Premier équipage arrivé en moins de 10 min : **90 % des cas**
- Premier équipage arrivé en moins de 12 min : **95 % des cas**

/ FIN DU RAPPORT FINAL /

LISTE DES FIGURES

- Figure 1** - Exemple de tableaux de bord de la LFB indiquant les incidents enregistrés
- Figure 2** - Liste des variables des jeux de données "Incidents"
- Figure 3** - Liste des variables des jeux de données "Interventions"
- Figure 4** - Mapping des catégories de variables pertinents pour notre modèle
- Figure 5** - Segmentation du temps d'intervention
- Figure 6** - Heatmap des incidents
- Figure 7** - Heatmap des mobilisations/interventions
- Figure 8** - Brainstorming de la data visualization
- Figure 9** - Evolution du temps d'intervention par année et par type d'incidents
- Figure 10** - Boxplot du temps d'intervention par année
- Figure 11** - Distribution des séquences de temps
- Figure 12** - Corrélation entre TravelTime, TurnoutTime et AttendanceTime (temps de déplacement, temps de mobilisation et temps d'arrivée)
- Figure 13** - Distribution du nombre d'interventions par arrondissement
- Figure 14** - Temps d'intervention moyen par zone géographique (Londres)
- Figure 15** - Carte interactive du nombre de pompes déployées des incidents
- Figure 16** - Evolution et volume des fausses alarmes par année
- Figure 17** - Répartition des types d'incidents
- Figure 18** - Catégories détaillées des incidents
- Figure 19** - Focus sur les actions issues des incidents de type 'special service'
- Figure 20** - Volume d'incidents en fonction des années (dataset = mobilisations)
- Figure 21** - Volume d'incidents par heure de la journée
- Figure 22** - Temps d'intervention par nombre de pompes déployées
- Figure 23** - Temps moyen d'intervention 1ère et 2ème équipe
- Figure 24** - Nombre de retards par motif de retard
- Figure 25** - Nombre de retards par motif de retard sans les "Not held up"
- Figure 26** - Données de la variable TravelTimeSeconds
- Figure 27** - Données de la variable TurnoutTimeSeconds
- Figure 28** - Tableau explicatif de la transformation des données du jeu "Incidents"
- Figure 29** - Tableau explicatif de la transformation des données du jeu "Mobilisations"
- Figure 30** - Heatmap des corrélations entre les colonnes numériques
- Figure 31** - Graphique en barres des 10 features les plus importantes
- Figure 32** - Graphique de comparaison des valeurs réelles vs prédites (Elastic Net)
- Figure 33** - Graphique de la distribution des erreurs (résidus) du modèle Elastic Net
- Figure 34** - Graphique d'analyse des résidus du modèle Elastic Net
- Figure 35** - Graphique d'analyse des résidus cumulés du modèle Elastic Net

BIBLIOGRAPHIE

Données

Deux principaux jeux de données disponibles ici :

- Détails des incidents à partir de 2009 :
<https://data.london.gov.uk/dataset/london-fire-brigade-incident-records>
- Détails des interventions depuis 2009 :
<https://data.london.gov.uk/dataset/london-fire-brigade-mobilisation-record>

Documents

- <https://www.london-fire.gov.uk/media/6686/crmp-metrics-30-may.pdf>

Sites Internet

- <https://www.london-fire.gov.uk>
- https://fr.wikipedia.org/wiki/London_Fire_Brigade