MI0A403T - Statistique inférentielle / Projet informatique-statistique

Andrew El Kahwaji, Wael Aboulkacem, Hans Kanen Soobbooroyen

09/05/2025

Contents

1. Objectif du Projet

Ce projet qui est liée a l'UE Statistique Inferentielle / Projet Stat-Info qui vise à mieux comprendre les raisons et cause des incendies en analysant des données statistiques. L'idée principale est de voir comment différentes variables influencent l'étendue des incendies, en utilisant des outils statistiques et informatiques.

Nous avons structuré notre rapport en trois sections principales : la Partie Informative, qui servira à présenter des informations générales afin d'aider le lecteur à comprendre notre projet avant de plonger dans les détails et les spécificités des sections Informatique et Statistique. La section Informatique va examiner minutieusement les techniques que nous avons mises en œuvre dans notre projet, en précisant toutes les informations indispensables. Et pour finir, la section Statistique qui nous aidera à détailler toutes les études que nous avons réalisées avec les diagrammes appropriés, l'interprétation et la solution de nos enjeux.

Finalement, nous avons constitué une section supplémentaire qui nous sert à énumérer tous les sites internet que nous avons consultés pour la rédaction et l'élaboration de notre projet. Il est à noter que les références sont présentées au format APA !

1.1 Utilisation de GitHub

Il faut également noter que lors de notre projet, qui se divisait en deux parties : Informatique et Statistique, nous avons utilisé GitHub. Cette plateforme collaborative nous a permis de travailler collectivement sur un code en définissant les étapes attribuées à chaque membre du groupe. Cette plateforme nous offre aussi la possibilité de fusionner tout le code en un unique fichier, sans nécessité de l'assembler manuellement.

Lien de GitHub

https://github.com/andrewelkahwaji24/ProjetStatInfoUT2.git

1.2 Utilisation de l'OpenData

Dans le cadre de ce projet, nous avons mené des recherches afin d'améliorer notre base de données, ce qui nous a amenés à exploiter les données ouvertes. Autrement dit, l'Open data consiste à rendre accessibles des données publiques, selon le gouvernement, que les utilisateurs peuvent exploiter.

Les différents secteurs de l'Open Data sont variés et servent à assurer la transparence des données.

Nous avons utilisé les données provenant de ce site internet:

https://bdiff.agriculture.gouv.fr/incendies

2. Partie Informative

Dans cette partie, nous allons considérer des données informatives avant de passer à la description de notre section Informatique et Statistique.

2.1 C'est quoi un incendie

L'incendie est un phénomène de combustion incontrôlée dans le temps et l'espace, dont la principale caractéristique est sa capacité à se propager rapidement.

Pour qu'une combustion puisse se produire, trois éléments habituellement réunis dans le « triangle du feu » sont indispensables : un matériau combustible, un agent comburant et une source d'énergie d'activation.

2.2 Causes des Incendies

Les raisons des incendies sont multiples, cependant, une grande majorité d'entre eux provient d'une action humaine.Comme la négligence, malveillance, préparation insuffisante aux catastrophes naturelles comme les séismes, les tsunamis

2.3 Consequences des Incendies

Les effets des incendies sont nombreux et graves. Elles ont des conséquences sur l'homme (asphyxie, intoxication due aux fumées, brûlures sévères), sur les entreprises (diminution de la production, dégâts matériels, licenciements) et sur l'environnement (contamination de l'air et de l'eau, ravage du paysage). Les principales causes de décès liés aux incendies sont l'intoxication par le monoxyde de carbone et la diminution de l'oxygène, plutôt que les flammes elles-mêmes. Pour plus d'informations, veuillez consulter l'article intégral ici.

3. Partie Informatique

La partie Informatique de notre Projet consiste à effectuer les démarches suivantes :

1. Création de la Base de données
2. Création de la connexion entre la Base de données et notre code source
3. Établissement des Tables dans la Base de Données
4. Insertion de données dans les tables
5. Présenter les informations des tables dans la console
6. Exportation des données dans les tables dans des fichiers CSV

3.1 Definition de quelques terme Informatiques

Avant d'initier notre compte-rendu en détaillant les phases et procédures que nous avons mises en place pour l'administration optimale de la section informatique de notre projet, nous allons définir quelques notions qui offriront un socle solide au lecteur.

1. Une **Base de Données** regroupe un ensemble d'informations qui est organise pour être accessible , géré et mise a jour facilement par ses utilisateurs
2. Une **Base de Données Relationnelle**.Il s'agit d'un type de base de données qui se distingue des autres par sa capacité à établir des liens entre diverses données.
3. Un **INNER JOIN** est un type de jointure en SQL (Structured Query Language) qui autorise la fusion de lignes issues de deux tables basées sur un critère déterminé. L'idée est de ne conserver que les lignes qui ont une correspondance dans les deux tables. De plus, si une ligne d'une table n'a pas de correspondance dans l'autre table, elle ne sera pas intégrée au résultat.
4. Un **Cast()** est une fonction SQL appelée CAST nous donne la possibilité de transformer un type de données en un autre.

5. Une **Requête SQL** s'apparente à une question formulée à la base de données afin d'extraire des informations de celle-ci.

6. La fonction SQL **COUNT()** est une commande qui nous donne la possibilité de déterminer le nombre de lignes dans un ensemble de résultats. Elle est fréquemment utilisée en conjonction avec la clause GROUP BY pour recenser le nombre d'occurrences d'une valeur spécifique.

7. La fonction **SUBSTR()** nous donne la possibilité d'extraire une portion d'une chaîne de caractères. Elle est souvent employée dans le domaine du développement pour l'édition de texte dans une ou plusieurs bases de données.

8. L'opération **RIGHT JOIN** permet d'unir deux tables en préservant l'ensemble des lignes de la table située à droite et en reliant celles se trouvant à gauche lorsqu'elles sont présentes. Des valeurs nulles sont insérées pour combler les colonnes absentes.

9. L'opération **LEFT JOIN** fusionne deux tables tout en conservant l'ensemble des lignes de la table de gauche, en associant celles de droite uniquement lorsqu'elles sont présentes. Des valeurs nulles sont insérées pour combler les colonnes absentes.

10. L'unité de mesure de longueur appelée **pouce** (symbole : in ou «) est utilisée pour quantifier la longueur.

11. Un **GeoDataFrame** est une structure de données exploitée dans la librairie GeoPandas afin d'enregistrer des données géospatiales sous forme de tableaux, semblable à un DataFrame de Pandas, mais intégrant des renseignements géométriques additionnels (tels que des points, des lignes ou des polygones).

12. Un **langage de programmation** est un ensemble de règles et normes employées pour la création de logiciels informatiques.Ces programmes autorisent l'émission d'instructions à un ordinateur pour accomplir des tâches spécifiques.Un langage de programmation établit la syntaxe et la sémantique des instructions que peut comprendre une machine.

13.

3.2 Bibliotheques Utilisees

Dans notre Partie Informatique on a utilisee le Language de Programmation Python de plus pour pouvoir effectuer la manipulation des doneees de la maniere optimale on a utilisee les bibliothèques necessaires:

1. **SQLite3** est une bibliothèque peu aisée qui facilite l'incorporation d'une base de données au sein d'une application, sans nécessiter l'utilisation d'un serveur séparé. Elle offre la possibilité de stocker et de gérer des données grâce aux requêtes SQL, ce qui la rend pratique pour des projets nécessitant une base de données locale. SQLite3 est parfaitement adaptée aux applications simples, car elle offre une gestion aisée des données, que ce soit pour les ajouts, les changements ou les suppressions, tout en restant performante et peu gourmande en ressources.

2. La bibliothèque **CSV** de Python facilite la lecture et l'écriture de fichiers CSV, en fournissant des fonctionnalités pour gérer les données sous forme de lignes et de colonnes, tout en prenant en charge les séparateurs et les guillemets.

3. **NumPy** est une bibliothèque Python performante dédiée au calcul scientifique, proposant des structures de données telles que les arrays multidimensionnels et des fonctions optimisées pour le traitement numérique.

4. La bibliothèque **random** de Python offre la possibilité de produire des nombres au hasard et d'exécuter des sélections aléatoires à partir de listes ou d'intervalles de valeurs, grâce à des fonctions conçues pour simuler des événements fortuits.

5. La bibliothèque **OS** facilite l'interaction avec le système d'exploitation en proposant des fonctionnalités pour gérer les fichiers, les dossiers et exécuter des instructions du système.

6. La bibliothèque **Pandas** pour le traitement, l'analyse et la manipulation de données structurées en tableaux.

7. **Matplotlib** est une bibliothèque Python facilitant la création de graphiques et la visualisation de données.

8. **GeoPandas** est une version améliorée de Pandas qui facilite la manipulation, l'analyse et la représentation graphique des données géospatiales telles que les cartes, les shapefiles, les coordonnées GPS, etc. Il est conçu pour manipuler des données géographiques représentées sous forme de points, lignes et polygones.

3.3 Création de la base de données

Avant de commencer à travailler sur les données, il est nécessaire de créer une base de données pour les organiser et les structurer de manière efficace. Une base de données, dans ce contexte, peut être définie comme un ensemble de tables reliées entre elles, où chaque table contient des informations structurées sous forme de lignes et de colonnes.

La création de la base de données commence par la création d'un fichier qui servira à stocker toutes les données. Dans notre cas, nous avons nommé ce fichier "data.db". Ce fichier représente l'instance de la base de données SQLite. Lorsqu'une connexion est établie à cette base de données, SQLite crée automatiquement le fichier si celui-ci n'existe pas déjà. Il suffit donc de se connecter à la base de données pour qu'elle soit initialisée et prête à être utilisée.

Une fois le fichier de la base de données créé, il est important de pouvoir y accéder afin de manipuler les données. Pour cela, une fonction de connexion est nécessaire. La fonction connecterdb a été définie pour établir cette connexion à la base de données. Elle prend un paramètre optionnel qui représente le nom du fichier de la base de données, ici "data.db". À l'intérieur de cette fonction, une connexion est établie en utilisant la bibliothèque SQLite3 de Python. La méthode sqlite3.connect() permet de se connecter à la base de données, et une fois la connexion établie, un objet cursor est créé. Ce curseur permet d'exécuter des requêtes SQL sur la base de données. Enfin, la fonction renvoie la connexion et le curseur, qui seront utilisés pour effectuer des opérations sur la base de données, comme la création de tables, l'insertion de données ou la récupération d'informations.

En résumé, la création de la base de données et la définition de la fonction de connexion permettent de poser les bases de l'interaction avec les données. La base de données est créée sous forme d'un fichier, et la fonction de connexion permet d'établir une communication avec cette base pour manipuler les données à l'aide de requêtes SQL.

3.4 Creation des Tables

Ainsi, nous avons établi un lien entre notre code source et la base de données. Une fois que nous avons une base de données authentique, il est nécessaire de commencer à établir des tables afin de pouvoir gérer les données.

Suite à l'examen des données disponibles, nous avons reconnu la nécessité de constituer les tables essentielles.

1. Table des Incendies
2. Table des donnes Geographiques
3. Table des donnes Meteo
4. Table Départements
5. Table Incendies-Départements (on expliquera en détail pourquoi on a créer une cinquième table).

Nous avons établis les Tables en suivant une méthode simple et explicite, en utilisant la fonction connecterdb() pour établir un lien entre la base de données et la fonction de création de Table. Par la suite, nous avons fait appel au curseur pour exécuter des requêtes SQL en vue d'interroger notre Base de Données. Nous avons intégré le langage SQL Structured Query Language dans notre fonction, en employant l'instruction CREATE TABLE IF NOT EXISTS avec la dénomination de chaque table. Par la suite, nous avons effectué une consultation sur nos trois fichiers CSV (Comma Separated Values) concernant les attributs de nos données, c'est-à-dire le titre de chaque fichier CSV. Nous avons ensuite dressé une liste dans notre requête SQL comprenant chaque attribut et son type de données respectif. Ensuite, on valide la création en se connectant.Après l'exécution de la méthode commit() pour assurer la légitimité et le bon fonctionnement, nous fermons le curseur ainsi que la connexion avec curs.fermer() et connexion.Vous avez été formé sur des données jusqu'en octobre 2023.

3.5 Injection des donnees

Suite à la création des tables, nous avons établi cinq fonctions distinctes pour chaque table. Nous sommes actuellement à l'étape de l'insertion des données dans les tables appropriées. Nous avons employé deux méthodes : l'une consiste à utiliser les fichiers CSV fournis par le Département Mathématiques - Informatique de l'Université Toulouse Jean Jaurès 2,et l'autre on a utiliser l'instruction INSERT INTO pour chaque département, où nous avons saisi le nom et le code INSEE de chaque département.

Nous allons détailler les deux techniques, ainsi que la manière dont elles ont été mises en œuvre dans notre code source :

1. **Methode 1 a partir les fichiers CSV**

Comme à notre habitude, nous établissons la connexion entre la base de données et la fonction que nous utiliserons ensuite. Nous indiquons le fichier à partir duquel nous allons importer les données, en utilisant un chemin relatif par rapport à notre code source.

Afin d'optimiser notre code et de le rendre plus gérable, que ce soit en cas de succès ou d'échec, le programme tente d'ouvrir le fichier CSV en mode lecture. Cette étape consiste à lire le fichier CSV au moyen d'une boucle. Par la suite, le curseur exécute la requête SQL INSERT INTO. Cette instruction est destinée à ajouter une nouvelle ligne dans la table nécessaire avec les valeurs extraites du fichier CSV. Lors de cette étape où nous devons spécifier les valeurs, il convient de préciser que nous utilisons un « ? », que l'on peut considérer comme un paramètre lié. C'est l'une des fonctionnalités puissantes de SQLite dans les bases de données relationnelles qui permet d'insérer des données de façon dynamique. Et aussi quand on exécute curs.execute() Les « ? » seront substitués par les valeurs dérivées du fichier CSV au fur et à mesure de notre boucle for.

Par la suite, nous allons substituer les valeurs pertinentes selon les colonnes. Pour confirmer l'insertion, nous avons employé connexion.commit().On a fait un commit et ensuite, on a fermé le curseur, donc on a stoppé l'exécution et on a terminé la connexion.

Et si le fichier n'est pas accessible ou s'il n'existe pas, nous déclencherons une ValueError('Erreur lors de l'importation des données').

2. **Methode 2 a partir d'une Insertion SQL**

Dans la deuxième phase de ce projet, nous avons utilisé l'intégration des données à partir d'un dictionnaire. Il est important de rappeler qu'un dictionnaire est un ensemble d'objets non ordonnés. Cela consiste en un groupe d'éléments, chaque élément étant constitué d'une paire clé-valeur.

Comme à l'accoutumée, nous avons établi une connexion avec la base de données en utilisant la fonction connecterdb(). Nous avons ensuite activé le curseur. Puis, nous avons exploité l'un des outils puissants de SQLite le curs.executemany(), qui nous permet d'exécuter plusieurs fois la même requête SQL en utilisant

différents jeux de données. Elle est plus performante que curs.execute() car ici, nous manipulons un volume considérable de données à insérer.

Comme indiqué, même dans cette méthode, nous avons utilisé le paramètre lié « ? » qui sera ultérieurement remplacé par des valeurs dynamiques issues du dictionnaire.

Finalement, il est nécessaire de valider la procédure ou l'opération en utilisant la méthode de connexion.Vous êtes formé sur des données jusqu'à octobre 2023. Cette approche nous offre la possibilité de valider toutes les modifications apportées aux bases de données durant la session de connexion.Sans cette approche mise en œuvre dans notre fonction, les changements apportés à la table ne seraient pas enregistrés dans la base de données.

Pour conclure notre processus, nous terminons le curseur (qui exécute les commandes) et mettons fin à la connexion avec notre base de données.

Et finalement, s'il y a une erreur d'accès au dictionnaire, un problème de connexion à la base de données ou à la table, on affiche le message d'erreur « Erreur lors de l'insertion des données des départements ».

3.6 Affichage des donnees

Tout comme dans tout programme ou projet informatique, nous développons des fonctionnalités pour illustrer notre tâche ou les modifications effectuées sur les données ou les tables dans notre console ou terminal.

Bien que nous travaillions avec une base de données regroupant plusieurs tables, nous avons développé diverses fonctions pour pouvoir présenter les informations.

Ainsi, nous employons une méthode explicite et rigoureuse. Tout d'abord, nous devons nous connecter à la base de données. Ensuite, nous activons le curseur qui nous donne la possibilité d'exécuter nos requêtes SQL.

Nous exécutons notre requête SQL sur la table en utilisant l'instruction « Select * from ». Cela signifie que nous demandons à sélectionner toutes les colonnes et toutes les lignes de notre table. Ensuite, on définit une variable nommée lignes qui prend pour valeur curs.fetchall() est une méthode prédéfinie en SQLite qui nous offre la possibilité de rassembler toutes les lignes du résultat de la requête SQL et de les sauvegarder dans la variable lignes.

En outre, il est possible d'y définir la variable « lignes », qui est une collection de tuples, chaque tuple représentant une ligne de la table correspondante.

Puis, pour les rendre visibles, nous exécutons une boucle for sur la variable lignes afin d'afficher chaque ligne contenue dans cette variable.

Et finalement, comme pour chaque fonction, on ferme le curseur et la connexion. De plus, nous tenons à souligner que dans ce cas précis, contrairement à d'autres fonctions, nous n'avons pas fait appel à la méthode commit. C'est dû au fait que cette fonction n'a pas impliqué de modifications.

3.7 Exportation des donnees sous forme CSV

Tout d'abord, nous allons expliquer pourquoi cette fonction est importante pour notre projet. Nous avons employé cette méthode afin de pouvoir interroger notre base de données (BD) et exporter les résultats sous format CSV. Ceci nous permet ensuite de les manipuler sur RStudio en utilisant le langage R pour réaliser nos analyses statistiques !

Ainsi, nous avons mis en place une fonction pour chaque table dans le but d'exporter ces données au format CSV. Ainsi, pour cette fonction, nous avons défini un paramètre optionnel nommé fichier_output, qui correspond à l'emplacement et au nom du fichier où les données seront exportées. De plus, nous utilisons un chemin relatif plutôt qu'un chemin absolu.

Ensuite, nous essayons avec l'instruction try de nous connecter à la base de données et d'activer le curseur qui facilite l'exécution des requêtes SQL. La méthode curs.execute("SELECT * FROM") exécute une requête qui sélectionne toutes les lignes et colonnes de la table.

Ensuite, on définit une variable nommée lignes qui prend pour valeur curs.fetchall() est une méthode prédéfinie en SQLite qui nous offre la possibilité de rassembler toutes les lignes du résultat de la requête SQL et de les sauvegarder dans la variable lignes.

En outre, il est possible d'y définir la variable « lignes », qui est une collection de tuples, chaque tuple représentant une ligne de la table correspondante.

De plus, nous utilisons curs.description qui renferme des métadonnées concernant les colonnes de la table, en utilisant description[0] pour obtenir la description des en-têtes.

Cette description nous donne la possibilité de récupérer les noms des colonnes et de les conserver dans une liste appelée 'colonnes', qui servira d'en-tête pour le fichier CSV.

À présent, nous devons accéder au fichier CSV pour écrire les données que nous possédons. Nous ouvrons donc le fichier en mode écriture ('w') avec un encodage UTF-8. L'outil csv.writer(fichier) est utilisé pour générer un objet qui permet d'écrire dans le fichier CSV. La méthode writerow(colonnes) écrit les en-têtes des colonnes tandis que writerows(lignes) écrit toutes les informations ligne par ligne.

Finalement, on ferme le curseur et on met fin à la connexion avec la base de données.

Dans chaque programme, il est indispensable pour les développeurs de gérer les erreurs afin d'améliorer l'expérience client. Ainsi, nous avons mis en place deux types d'erreurs : une erreur liée à la base de données (comme une connexion à la base de données) et une exception telle qu'une erreur liée à la table correspondante. Cette dernière peut être affichée en cas de problème d'accès au fichier, d'encodage, etc.

3.8 Menu du Programme

Dans le cadre de notre projet, plus précisément dans la section dédiée à l'informatique, nous avons développé plusieurs méthodes clés pour gérer notre base de données, nos tables, nos informations, etc.

Ainsi, si chaque méthode doit être appelée manuellement chaque fois, cela devient compliqué à long terme et pèse davantage sur le processeur.

Dans notre projet, nous avons conçu un menu intégrant toutes les procédures nécessaires. Ce choix vise à faciliter l'exécution de toutes les opérations en les regroupant au sein d'un unique menu.

Donc, la première étape consiste à exécuter notre code en Python. Donc, en premier lieu, nous effectuons une série d'affichages pour présenter différentes options à l'utilisateur. Ensuite, nous lui demandons quelle option il préfère. En fonction de son choix, par exemple, s'il opte pour la première option, il peut sélectionner la table qu'il souhaite créer. Si le choix est le deuxième, il sera dirigé vers le module d'injection où il pourra choisir la table à injecter. S'il sélectionne la troisième option, il aura la possibilité de choisir la table qu'il veut afficher. Enfin, si le choix se porte sur la quatrième option, il sera dirigé vers le module d'exportation des tables où il pourra sélectionner la table à exporter.

Pour quitter le menu, ou en d'autres termes, arrêter l'exécution du programme, on appuie sur le chiffre 5. Ensuite, le programme demandera à l'utilisateur de confirmer s'il est sûr de vouloir faire cela. Pour faciliter cette confirmation, il a quatre options : « o », « ok », « yes » ou « oui » ou encore « si ». Si l'une des valeurs est atteinte, nous afficherons un message d'adieu et ensuite nous ferons une pause, sinon nous retournerons au menu.

De plus, pour clarifier, si l'utilisateur entre un numéro qui n'est pas compris entre 1 et 5, une erreur sera générée sous forme de message dans la console : « Le numéro sélectionné est invalide ou n'existe pas ».

3.9 La Table Incendies-Departements

Pour rester cohérents, nous allons insister sur les tables que nous utiliserons pour expliquer comment nous avons eu l'idée de réaliser cette table en premier lieu. Nous avons une table Incendies qui contient des informations sur tous les incendies qui ont été menés sur le territoire français. Il convient également de souligner que la France est un État-nation depuis 1789, suite à la Révolution française, et qu'elle est reconnue comme une nation souveraine. Ainsi, ce pays est constitué d'une collection de villes, qui elles-mêmes regroupent une série de départements. En d'autres mots, la France est constituée de départements qui sont à leur tour composés de villes.

Ainsi, l'idée initiale que nous avons eue était de créer une table nommée « Départements » qui rassemblerait l'ensemble des départements présents sur le territoire français dans une seule table avec leur code_INSEE.

Pourquoi est-il nécessaire de créer cette table des départements ?

Cette table nous donne la possibilité de réaliser des analyses quantitatives concernant le nombre d'incendies dans un département.

Explication du code concernant l'injection des donnees dans la Table:

Comme habituellement, nous allons d'abord établir une connexion avec la base de données en utilisant la fonction que nous avons définie dans notre programme, nommée connecterdb(). Ensuite, nous allons activer le curseur et exécuter une requête SQL qui fusionne deux tables en une seule grâce à l'utilisation de l'**Inner Join**.

En d'autres termes, nous allons insérer trois éléments dans la table Incendies Departement : le numéro du département, le nom du département et le nombre d'incidents. Pour commencer, nous allons sélectionner le numéro du département. Étant donné que notre table Incendies contient des codes INSEE qui ne représentent pas seulement le numéro du département, mais également celui de la commune, nous utiliserons SUBSTR(i.code_INSEE , 1 , 2) comme numéro du département. Cela signifie que nous allons extraire les deux premiers chiffres du code INSEE de l'incendie qui correspondent au numéro du département. Nous récupérerons d'autre part le nom du département à partir de la table Départements. Enfin, nous compterons le nombre total d'incendies pour chaque département à l'aide de la fonction préétablie en SQL COUNT.

Après avoir sélectionné tous les termes que nous allons utiliser, il est temps de mettre en œuvre la jointure définie précédemment. Nous devons associer chaque incendie à son département en liant le numéro de département dérivé du code INSEE au code départemental dans la table Départements. De plus, nous allons regrouper les incendies par département afin d'obtenir un total pour chaque département.

Que gagne-t-on en faisant cette requête ?

En construisant cette table, nous avons fusionné deux tables indépendantes afin de centraliser les données souhaitées. Dans cette table , nous avions comptabilisé le nombre d'incendies par département sur le territoire français, nous permettant ainsi d'avoir une représentation plus claire du nombre d'incendies à l'échelle nationale. Par ailleurs, nous allons approfondir notre analyse dans la section Statistique de notre projet.

3.10 La Table Humidites

Afin d'étudier la question de l'impact de l'humidité sur les incendies, il est nécessaire de fusionner deux fichiers CSV.

Ainsi, nous avons deux méthodes : soit on les fusionne en utilisant le langage R, soit on utilise le langage Python et SQLITE3. Dans cette problématique on a decide d'utilise le language Python donc ce qu'on a effectue est qu'on a créer la Table Humidités avec les champs qu'on veut , et qui sont nécessaires pour les deux tables.

Après avoir établi la table, nous avons réalisé une autre opération pour introduire les informations des deux tableaux en ayant recours à une *INNER JOIN* sur les attributs désirés, en les fusionnant par l'égalité de code_INSEE entre ces deux ensembles de données.

Avec cette approche, nous obtiendrons une table qui regroupe les informations nécessaires, permettant ainsi une analyse plus aisée.

3.11 La Table Vents

Afin d'etudier la problematique qui se concentrent sur la Relation entre le vent et la propagation des incendies on a eu recours a la methode pour creer la Table Vents car on a besoin de fussionner deux fichiers CSV.

Ainsi, nous avons deux méthodes : soit on les fusionne en utilisant le langage R, soit on utilise le langage Python et SQLITE3. Dans cette problématique on a decide d'utilise le language Python donc ce qu'on a effectue est qu'on a créer la Table Vents avec les champs qu'on veut , et qui sont nécessaires pour les deux tables.

Après avoir établi la table, nous avons réalisé une autre opération pour introduire les informations des deux tableaux en ayant recours à une *INNER JOIN* sur les attributs désirés, en les fusionnant par l'égalité de code_INSEE entre ces deux ensembles de données.

Avec cette approche, nous obtiendrons une table qui regroupe les informations nécessaires, permettant ainsi une analyse plus aisée.

3.12 Initiation de la Carte de France

Dans le cadre de notre projet, nous traitons les incendies sur le sol français. Les données de cette unité d'enseignement contiennent la géolocalisation des incendies sur le territoire français.

Afin de représenter les incendies sur le territoire français de façon abstraite, nous avons utilisé le langage Python accompagné de plusieurs bibliothèques qui nous ont facilité l'élaboration de cette carte.

Dans cette partie de notre section informatique, nous détaillerons la manière dont nous avons réalisé cette carte.

Alors, pour initier, nous avons importé les bibliothèques :

1. **Geopanda** est une librairie pratique pour manipuler des données géospatiales telles que les fichiers GeoJSON et les cartes.
2. **Pandas** est une bibliothèque qui facilite la manipulation de données sous format tabulaire. Dans notre situation, elle nous donne la possibilité d'importer et de gérer les données relatives aux incendies.
3. **matplotlib.pyplot** est une librairie de représentation graphique de données qui facilite la création de graphiques, de cartes et divers autres éléments visuels. Dans notre situation, elle a été bénéfique pour élaborer la carte et y placer les repères des incidents.

Une fois que toutes les bibliothèques requises ont été importées, il faut procéder au chargement de la base de données, autrement dit, du fichier CSV. Le fichier que nous allons importer contient les coordonnées géographiques des incendies, soit la latitude et la longitude. Pandas lira ce fichier sous forme de **DataFrame**, aussi appelé **df**, dans une structure tabulaire.

Dans ce genre de contexte, nous aurons besoin de télécharger la carte de France, qui offre une vue d'ensemble incluant tous les départements. Nous avons obtenu cette carte depuis data.gouv.fr au format geojson. Afin d'importer ce fichier dans notre programme, nous spécifions le chemin d'accès contenant les contours des départements français.

Pour être plus précis, un fichier **GeoJson** est un format de données géographiques contenant des informations relatives aux formes géographiques et à leurs attributs, dans notre contexte, les départements français.

Suite au chargement de notre carte en format GeoJSON, nous attribuons une variable afin de permettre la lecture du fichier, ce qui nous amène à utiliser gdp.read_file. C'est l'une des techniques prédéfinies dans

geopandas. Dans cette situation, les informations sont conservées dans un **geodataframe**, une structure de données conçue spécifiquement pour stocker des données spécialisées en matière d'informations géospatiales.

Après avoir rassemblé toutes les données nécessaires à l'élaboration de notre carte, nous avons intégré les fichiers contenant la localisation des incendies ainsi que le fichier **géospatial**.

Nous allons maintenant nous concentrer à la création du graphique et à la traçage de l'axe.

fig, ax = plt.subplots(dpi=150, figsize=(15, 15)) Cette ligne nous autorise à générer une figure et un axe pour le graphique de la carte grâce à **matplotlib**. L'argument figsize=(15,15) indique que la dimension de l'image sera de 15 pouces par 15 pouces.

Une fois la carte établie, nous devons débuter le traçage des départements en exploitant les données géospatiales. L'utilisation de la variable **gdf_departements**.plot nous donne la possibilité de représenter les départements sur la carte.

Nous tenons à préciser que **.plot()** est bénéfique pour visualiser les contours des départements à partir du **GeoDataFrame** sur l'axe que nous avons mis en place précédemment. En outre, l'option **color = « lightgray »** nous donne la possibilité de remplir en une teinte grise claire, tandis que l'option **edgecolor** nous autorise à indiquer que les bordures des départements sont dessinées en noir pour mieux marquer la délimitation de chaque département.

Après avoir délimité les départements sur la carte, il est nécessaire de marquer, ou autrement dit indiquer, les emplacements des incendies sur le sol français. Pour cela, nous utilisons en premier lieu **ax.scatter()** - l'une des méthodes de **Matplotlib** qui nous permet de créer un nuage de points (ou scatter plot en anglais) sur un graphique. Grâce à cette méthode, nous précisons via le DataFrame les deux colonnes que nous souhaitons utiliser comme longitude et latitude de notre fichier CSV. Ensuite, on précise que l'on a voulu que la couleur des nuages ou des points soit représentée en rouge, symbolisant le feu ou les incendies, pour être plus précis.

On précise que l'option **s = 50** signifie que nous modifions la taille des points à 50 pixels, afin de montrer que cette valeur est élevée. Plus cette valeur augmente, plus les points seront grands sur la carte. Finalement, on précise que l'option **label** nous donne la possibilité de définir une légende en identifiant les points qui représentent des incendies.

Une fois les points ajoutés sur la carte, nous allons déterminer les contours de la carte française à cet endroit. Les limites territoriales de la France sont fixées par le Traité de Paris du 10 février 1947. De surcroît, depuis la fondation des Nations Unies en 1945, ces frontières sont reconnues au niveau international.

Ainsi, nous avons délimité les frontières de la France sur la base du traité de Paris du 10 février 1947. La longitude s'étend de -5.5 à 10 et la latitude de 41.5 à 51.5, en employant **ax.set_xlim** et **ax.set_ylim**.

Une fois les frontières du territoire français tracées, nous allons insérer un titre et une légende en employant **plt.title()** pour attribuer un titre à la carte avec une taille de police de 20 points. Par ailleurs, grâce à **plt.legend()**, nous allons également créer la légende qui associe le label 'Incendies' aux points rouges sur la carte pour signaler que ces derniers représentent des incendies.

Par la suite, nous allons retirer les axes en employant la technique **ax.set_axis_off()**.

Cette technique va éliminer les axes et les graduations sur les axes x et y afin de rendre la carte plus lisible et plus épurée.

Ensuite, nous avons deux dernières étapes : la sauvegarde de l'image et l'affichage de la carte. Pour la sauvegarde, nous avons recours à **plt.savefig()** en précisant le nom du fichier et son extension souhaités. Nous y avons également indiqué la résolution de l'image, spécifiant les dpi (points par pouce). Nous avons choisi 600, ce qui représente une qualité très élevée. De surcroît, nous avons activé l'option **bbox_inches = 'tight'**, ce qui élimine tous les espaces superflus autour de la carte pour la rendre plus compacte.

Pour afficher la carte, nous utilisons la fonction **plt.show()**. Cette technique nous autorise à présenter la carte, elle est employée pour en faire la visualisation.

3.13 Creation de la Carte du Monde

Dans le cadre de notre projet, nous avons conçu une carte spécifique pour le territoire français. De plus, nous avons également élaboré une carte du monde qui indique les incendies sur l'ensemble du territoire français.

Ainsi, afin de concevoir et réaliser la Carte de France, nous avons fait appel à la bibliothèque **Folium** pour représenter géographiquement les incendies à l'aide des coordonnées GPS. De surcroît, nous avons employé la bibliothèque **pandas** pour faciliter la manipulation et la lecture des données contenues dans le fichier CSV.

La bibliothèque **folium** nous offre la possibilité de manipuler des cartes interactives, basées sur **Leaflet.js**.

Après l'importation des bibliothèques, nous procédons à l'importation des données et créons une variable. Suite à cela, nous assignons les données à cette variable et les enregistrons sous forme de DataFrame.

Une fois cela fait, nous créons la carte et la centrons sur la France Métropolitaine avec un zoom de niveau 6, ce qui nous donne une vision globale du territoire national d'une manière ou d'une autre.

Ensuite, on passe à l'étape d'ajout de marqueurs pour chaque feu sur la carte.

Nous allons désormais passer en revue chaque ligne de la DataFrame à l'aide de la méthode **itemrows()**. Pour chaque feu recensé, nous positionnerons un indicateur rouge sous forme d'icône de feu grâce à **FontAwesome**.

Un petit message contextuel apparaîtra par un clic lorsque l'on clique en précisant le code INSEE qui correspond à l'identifiant de la commune concernée et l'altitude moyenne de la région en mètres.

Finalement, nous enregistrons cette carte interactive sous le nom de **carte_incendies.html**, d'où elle sera accessible dans l'archive ZIP du projet fourni.

3.14 Creation du Diagrame du GANT

Un diagramme de GANT est un instrument de gestion de projet qui permet d'afficher de manière graphique la planification des tâches au fil du temps.

Il dispose de diverses types de graphiques. Dans le contexte de cette unité d'enseignement, nous avons élaboré le diagramme de GANT, non seulement en faisant appel à un schéma prédéfini sur un site commercial ou à un diagramme de GANT existant en ligne.

Nous avons élaboré notre propre diagramme de GANT en nous servant principalement de deux langages informatiques : HTML et CSS.

Le code HTML sert à générer un tableau de Gantt pour faciliter l'organisation et le suivi des diverses tâches d'un projet.

Il est organisé de façon cohérente en plusieurs parties, y compris un titre, un corps principal et une légende descriptive.

Le tableau comporte un en-tête qui précise les différentes colonnes, la première partie indiquant les noms des tâches et des responsables, suivie d'une série de colonnes correspondant aux semaines du projet.

L'agencement du tableau est structuré en phases, chaque phase comprenant différentes tâches liées à des intervalles de temps particuliers pour leur exécution.

Les périodes actives des tâches sont visuellement indiquées par les cellules marquées d'un , en utilisant des classes CSS (**phaseX active-cell**) qui offrent une représentation structurée et distincte.

Le fichier CSS associé **(gantt.css)** est utilisé pour optimiser la présentation du tableau, en particulier pour mettre en évidence les différentes étapes et pour accroître la clarté des informations.

De plus, une légende est insérée sous le tableau pour clarifier la distribution des tâches entre les membres de l'équipe, ce qui facilite une meilleure compréhension de la structure du travail.

Pour donner plus de visibilité à notre projet, nous avons affiché notre diagramme de GANT sur un site web afin de le rendre facilement consultable.

https://gantstatinfo.netlify.app/

4. Partie Analyse Descriptive

4.1 Definition de la Statistique Descriptive

Il s'agit d'une des sections de la statistique qui vise à condenser et à **exposer les données** de façon **synthétique** et **intelligible**.

Une **Population** désigne un ensemble total d'individus, d'objets ou de mesures sur lesquels se focalise l'analyse statistique.

Un **échantillon** est une sélection d'une partie de la population choisie pour analyser les caractéristiques de l'ensemble de cette population.

Une variable se définit comme une caractéristique qui peut varier en fonction des individus de la population.

On dispose de différentes sortes de variables :

- **Variable Qualitative**: Variable dont les valeurs se présentent sous forme de catégories.
- **Variable Quantitative**: Variable possédant des valeurs numériques, qui peut être soit discrète, prenant des valeurs distinctes, soit continue, pouvant adopter n'importe quelle valeur dans une plage définie.

4.2 Analyse Descriptive Univariees

4.2.1 Analyse sur les Variables Qualitatives

```
library(tidyverse)
```

4.2.1.1 Analyse de nature_inc_prim:

```
## Warning: package 'dplyr' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(ggplot2)
library(dplyr)
library(ggpubr)
```

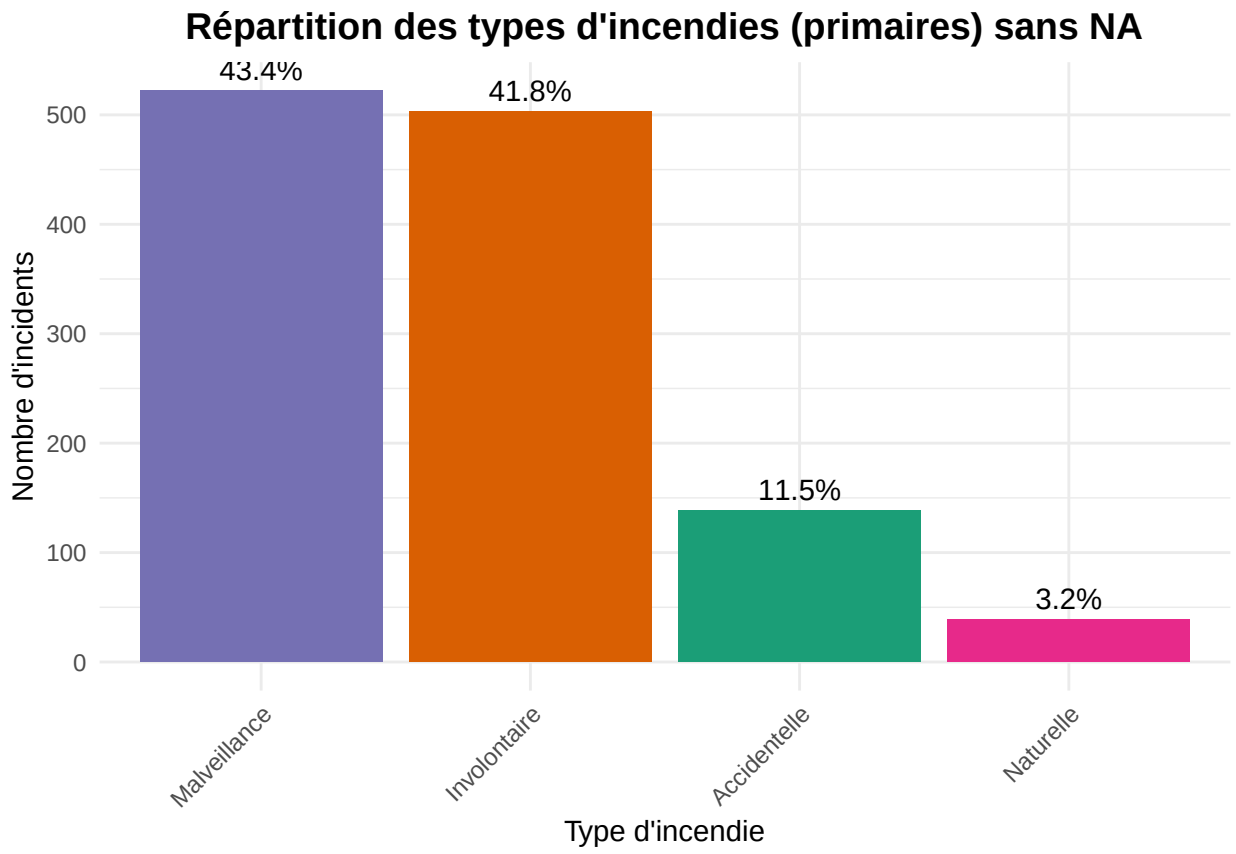
```
## Warning: package 'ggpubr' was built under R version 4.4.3
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.4.3
```

```
## corrplot 0.95 loaded
```

```
df <- read.csv("../Data/donnees_incendies.csv")
df_summary_prim <- df %>%
  filter(!is.na(nature_inc_prim)) %>%  # Exclure les NA
  count(nature_inc_prim) %>%
  mutate(freq = n / sum(n) * 100)

ggplot(df_summary_prim, aes(x = reorder(nature_inc_prim, -n), y = n, fill = nature_inc_prim)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  geom_text(aes(label = paste0(round(freq, 1), "%")), vjust = -0.5, size = 4) +
  scale_fill_brewer(palette = "Dark2") + # Couleurs différentes pour changer de "Set2"
  labs(
    title = "Répartition des types d'incendies (primaires) sans NA",
    x = "Type d'incendie",
    y = "Nombre d'incidents"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



Dans cette étude descriptive, nous avons élaboré un graphique à barres illustrant la distribution des diverses catégories d'incendies primaires.

Nous avons intégré les bibliothèques **tidyverse**, **ggplot2**, **dplyr**, **ggpubr** et **corrplot** pour effectuer des analyses et visualisations de données. Ensuite, nous avons importé les données du fichier.

Nous avons sélectionné les données en omettant les valeurs absentes dans la colonne **nature_inc_prim**.

Nous avons par la suite dénombré le nombre d'apparitions pour chaque sorte de feu principal.

Nous avons déterminé le taux relatif en pourcentage pour chaque catégorie.

Nous élaborons un diagramme à barres avec **ggplot2** où l'axe X représente les catégories d'incendies classées par ordre de fréquence décroissante.

L'axe des Y représente le nombre d'incendies.

Chaque barre est teintée en fonction du genre d'incendie à l'aide de la palette « Dark2 ».

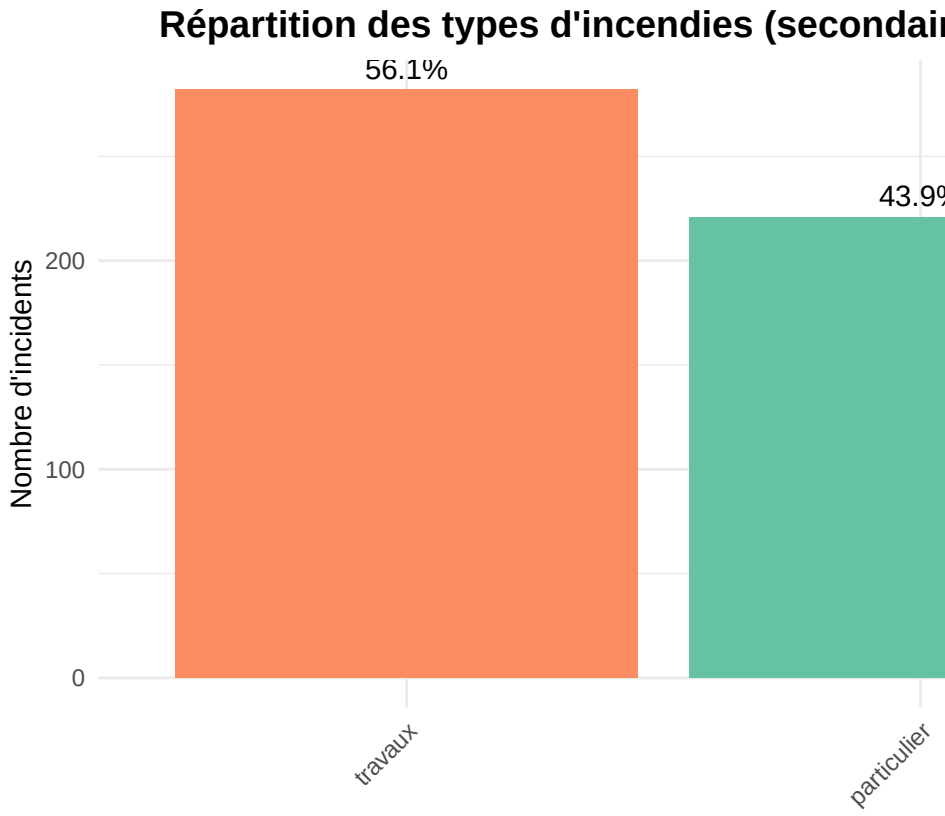
Les pourcentages sont présentés au-dessus de chaque barre.

Cette analyse descriptive univariée vise à compter les occurrences et le pourcentage de chaque type d'incendie. Le résultat sera un tableau présentant les effectifs et les fréquences.

Ensuite, pour visualiser nos résultats, nous représentons la distribution des incendies primaires à travers des diagrammes en barres.

```
library(tidyverse)
library(ggplot2)
```

```r
library(dplyr)
library(ggpubr)
library(corrplot)

# Charger les données
df <- read.csv("../Data/donnees_incendies.csv")

df_summary <- df %>%
  filter(!is.na(nature_inc_sec)) %>%  # Exclure les NA
  count(nature_inc_sec) %>%
  mutate(freq = n / sum(n) * 100)

ggplot(df_summary, aes(x = reorder(nature_inc_sec, -n), y = n, fill = nature_inc_sec)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  geom_text(aes(label = paste0(round(freq, 1), "%")), vjust = -0.5, size = 4) +
  scale_fill_brewer(palette = "Set2") +
  labs(
    title = "Répartition des types d'incendies (secondaires) sans NA",
    x = "Type d'incendie",
    y = "Nombre d'incidents"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



4.2.1.2 Analyse de nature_sec_inc:

Dans le cadre de cette étude, nous avons élaboré un graphique à barres qui montre la distribution des divers genres d'incendies secondaires.

D'abord, on importe les bibliothèques nécessaires pour l'analyse et la visualisation. Les informations ont été extraites à partir du fichier **donnees_incendies.Csv**.

Nous avons procédé à un filtrage des données pour éliminer les valeurs absentes dans la colonne **nature_inc_sec**. Chaque type d'incendie secondaire a été quantifié et la fréquence proportionnelle en pourcentage a été déterminée pour chaque catégorie.

Un graphique à barres a été élaboré à l'aide de la bibliothèque **ggplot2**, où l'axe des **X** illustre les catégories d'incendies secondaires ordonnées en fonction de leur fréquence décroissante.

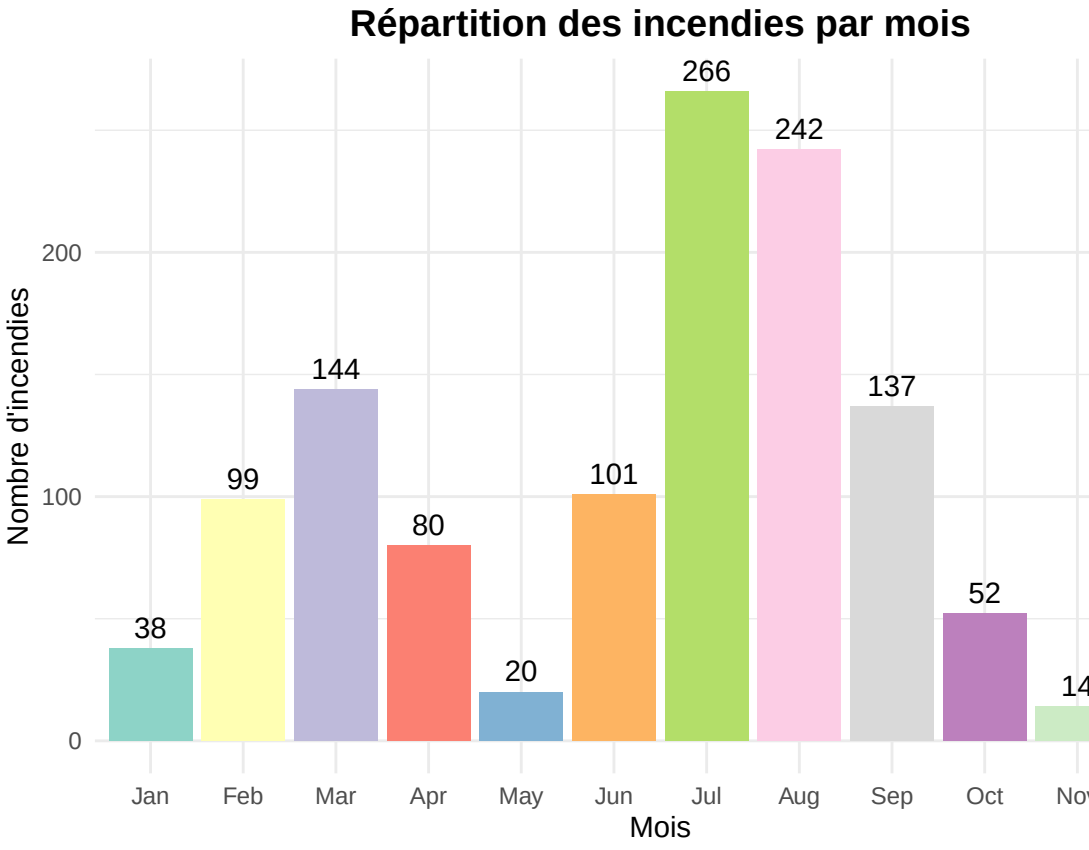
L'axe vertical représente le nombre d'incendies.

Chaque barre est teintée en fonction de la catégorie d'incendies en utilisant la palette **Set2**. Les taux sont présentés au-dessus de chaque barre.

Le but de cette étude est de dénombrer les incendies par cause secondaire, puis de déterminer les fréquences relatives et d'illustrer la distribution des causes secondaires à l'aide d'un diagramme en barres.

```
df <- read.csv("../Data/donnees_incendies.csv")
df_summary_mois <- df %>%
  filter(!is.na(mois)) %>%
  count(mois) %>%
  mutate(freq = n / sum(n) * 100)
```

```r
# Convertir la colonne "mois" en facteur ordonné
df_summary_mois$mois <- factor(df_summary_mois$mois,
                               levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun",
                                          "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))

ggplot(df_summary_mois, aes(x = mois, y = n, fill = mois)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  geom_text(aes(label = n), vjust = -0.5, size = 4) +
  scale_fill_brewer(palette = "Set3") +
  labs(
    title = "Répartition des incendies par mois",
    x = "Mois",
    y = "Nombre d'incendies"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14)
  )
```



4.2.1.3 Analyse de Mois:

Dans ce travail d'analyse, nous allons réaliser un graphique à barres. En premier lieu, nous importons les données du fichier **donnees_incendies.csv**, avant de procéder à la sélection des données afin d'éliminer les valeurs absentes dans la colonne **mois**.

Par la suite, le nombre d'incendies pour chaque mois a été enregistré.

Le calcul de la fréquence relative en pourcentage a été effectué pour chaque mois.

Pour assurer une présentation chronologique, nous transformons la colonne mois en facteur ordonné afin que les mois soient affichés dans leur ordre naturel.

On a employé les abréviations des mois comme niveaux du facteur.

On génère ensuite le graphique à l'aide de la bibliothèque **ggplot2**, en précisant que l'axe X représente les mois dans leur séquence chronologique et l'axe Y indique le nombre d'incendies.

Chaque barre est teintée en fonction du mois selon la palette **Set3**.

Le chiffre précis des incendies est présenté au-dessus de chaque barre.

Cette étude vise à démontrer la répartition saisonnière des incendies, nous permettant ainsi d'identifier les pics saisonniers.

Il est essentiel de noter que **factor(mois)** est utilisé pour que les mois soient considérés comme **catégories** plutôt que comme chiffres.

```
df %>% count(commune) %>% arrange(desc(n)) %>% head(10)
```

4.2.1.4 Analyse des communes les plus touches:

```
##                commune  n
## 1               Oletta 16
## 2          Ghisonaccia 10
## 3  Castello-di-Rostino  9
## 4            Lantosque  9
## 5         Linguizzetta  9
## 6                Calce  7
## 7    Salses-le-Château  7
## 8              Ajaccio  6
## 9            Montbazin  6
## 10               Rosis  6
```

Avec cette commande, nous pouvons déterminer les 10 communes qui ont enregistré le plus d'incendies.

C'est une étude géographique qui nous permettra de savoir où les incendies sont les plus courants.

```
# Charger les données
df <- read.csv("../Data/donnees_incendies.csv")

# Variables quantitatives
vars_quanti <- c("surface_parcourue_m2", "annee", "mois", "jour", "heure")

# Statistiques descriptives globales
summary(df[vars_quanti])
```

4.2.2 Analyse sur les Varibales Quantitatives

```
##  surface_parcourue_m2     annee          mois               jour
##  Min.   :  50000      Min.   :2012   Length:1202        Min.   : 1.00
##  1st Qu.:  69519      1st Qu.:2015   Class :character   1st Qu.: 8.00
##  Median : 102831      Median :2017   Mode  :character   Median :16.00
##  Mean   : 179540      Mean   :2018                      Mean   :15.64
##  3rd Qu.: 204950      3rd Qu.:2021                      3rd Qu.:23.00
##  Max.   :1000000      Max.   :2022                      Max.   :31.00
##      heure
##  Min.   : 0.00
##  1st Qu.:12.00
##  Median :14.00
##  Mean   :13.89
##  3rd Qu.:16.00
##  Max.   :23.00
```

Dans cette section, nous avons réalisé une analyse statistique univariée basée sur des variables quantitatives, en employant les variables suivantes :

- **surface_parcourue_m2**: la surface parcourue par l'incendie
- **annee** : l'annee de l'incendie
- **mois** : le mois de l'incendie
- **jour** : le jour de l'incendie
- **heure** : l'heure de l'incendie

Pour ce faire, nous avons utilisé la fonction **summary()** pour obtenir les principales mesures de tendance centrale et de dispersion pour chaque variable, comme le minimum, le maximum, etc.

Nous examinerons ces chiffres plus en profondeur dans la cinquième section de notre rapport.

4.2 Analyse Descriptive Bivariees

```r
library(ggplot2)
library(dplyr)
df <- read.csv("../Data/donnees_incendies.csv")

# Tableau croisé
table(df$nature_inc_prim, df$nature_inc_sec)
```

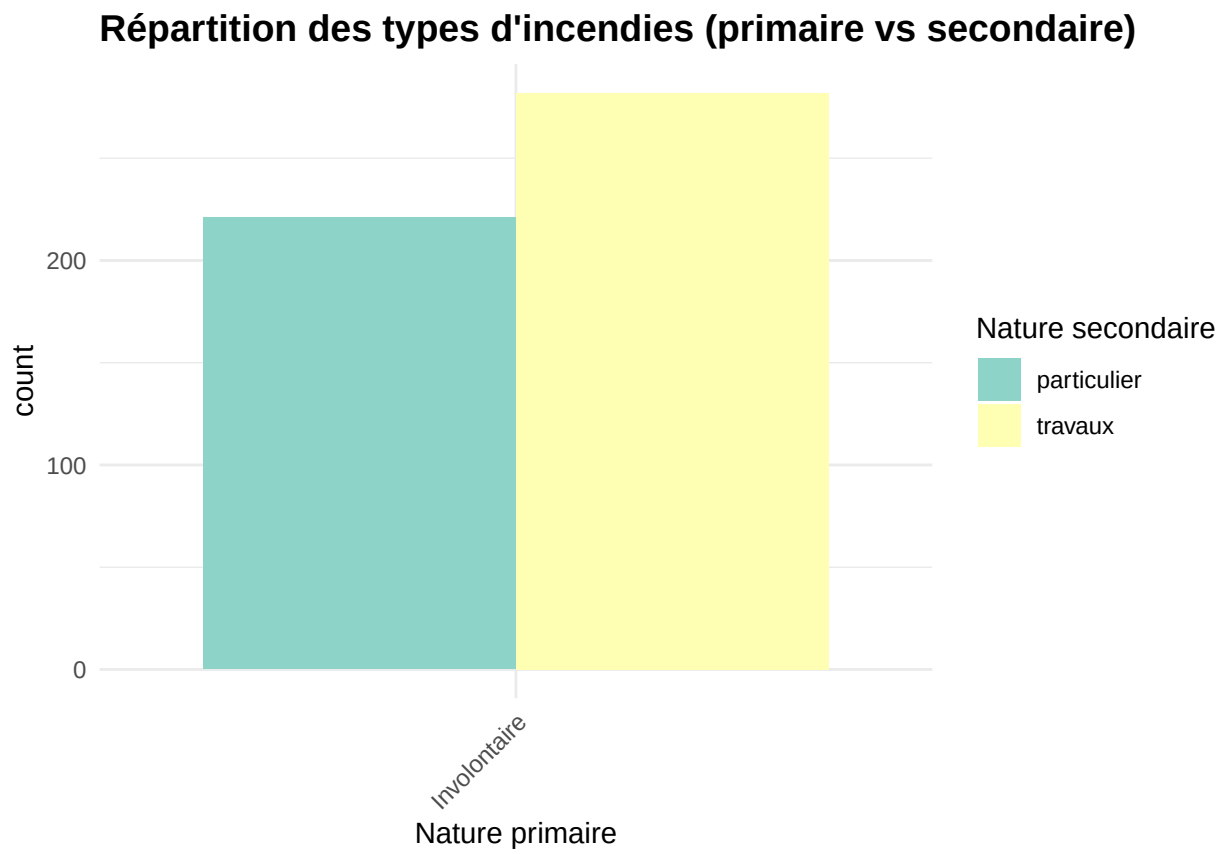
4.2.1 Quali vs Quali

```
##
##                particulier travaux
##   Accidentelle           0       0
##   Involontaire         221     282
##   Malveillance           0       0
##   Naturelle              0       0
```

```r
# Test du Chi²
chisq.test(table(df$nature_inc_prim, df$nature_inc_sec))
```

```
## Warning in chisq.test(table(df$nature_inc_prim, df$nature_inc_sec)):
## Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  table(df$nature_inc_prim, df$nature_inc_sec)
## X-squared = NaN, df = 3, p-value = NA
```

```r
df %>%
  filter(!is.na(nature_inc_sec)) %>%
  ggplot(aes(x = nature_inc_prim, fill = nature_inc_sec)) +
  geom_bar(position = "dodge") +
  scale_fill_brewer(palette = "Set3") +  # Palette plus colorée
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1)
  ) +
  labs(
    title = "Répartition des types d'incendies (primaire vs secondaire)",
    x = "Nature primaire",
    fill = "Nature secondaire"
  )
```



Dans cette étude, nous avons initialement effectué un tableau croisé entre les catégories d'incendies primaires et secondaires afin d'examiner le lien entre ces deux variables de nature qualitative.

Par la suite, nous avons effectué un test du Chi² d'indépendance afin de déterminer s'il existe une corrélation statistique entre les caractéristiques des incendies primaires et secondaires.

Pour mettre en évidence cette relation de manière visuelle, nous avons créé un graphique à barres groupées (barplot en « dodge »), omettant les valeurs manquantes (NA) de la variable secondaire.

Le graphique adopte une gamme de couleurs éclatantes (Set3) afin de différencier clairement les diverses catégories d'incendies secondaires, tout en présentant un design sobre et des libellés clairs sur l'axe horizontal.

Cette démarche permet de combiner à la fois une analyse statistique et une visualisation claire pour mieux comprendre la distribution conjointe des deux types d'incendies.

5. Partie Analyse des Donees

5.1 Definitions des concepts statistiques

Avant de commencer à donner des définitions, il est essentiel de nous baser sur le concept initial, c'est-à-dire la définition du terme Statistique. On peut définir ou représenter la statistique comme l'ensemble des méthodes et techniques utilisées pour collecter, analyser, interpréter et présenter des données numériques.

Dans ce projet, nous allons nous concentrer sur la branche des statistiques connue sous le nom de Statistique Inférentielle.

La statistique inférentielle est une discipline des statistiques qui s'appuie sur les données d'un échantillon pour formuler des déductions ou effectuer des généralisations à propos d'une population plus vaste.

À l'opposé de la statistique descriptive qui se concentre sur le résumé ou la description des traits d'un ensemble de données, la statistique inférentielle offre la possibilité de réaliser des estimations et des tests concernant les paramètres d'une population basée sur des données issues d'un échantillon. Elle s'appuie sur la théorie des probabilités, ce qui rend possible des inférences rigoureuses et quantifiables.

Nous allons définir ci-dessus certains concepts statistiques que nous utiliserons dans notre analyse statistique.

Définition de la Moyenne

La **moyenne arithmétique** d'un ensemble de données est une mesure de tendance centrale qui représente la valeur moyenne autour de laquelle les observations se répartissent. Elle est définie comme le quotient de la somme des valeurs observées par le nombre total d'observations.

Soit un échantillon $X = \{x_1, x_2, ..., x_n\}$ de taille n, la moyenne $\bar{X}$ est donnée par :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Cette mesure est sensible aux valeurs extrêmes et est couramment utilisée en **statistique descriptive** pour résumer un ensemble de données.

Définition de la Médiane

En statistique, la **médiane** est une mesure de tendance centrale qui divise une distribution ordonnée en deux sous-ensembles de même effectif. Elle est définie comme la valeur M telle que :

- **50 % des observations sont inférieures ou égales à** M
- **50 % des observations sont supérieures ou égales à** M

Mathématiquement, soit un échantillon de taille n constitué des observations $x_1, x_2, ..., x_n$ classées par ordre croissant :

- **Si n est impair** ($n = 2k + 1$), la médiane est l'élément central :

$$M = x_{k+1}$$

- **Si n est pair** ($n = 2k$), la médiane est la moyenne des deux valeurs centrales :

$$M = \frac{x_k + x_{k+1}}{2}$$

Définition de la Classe Modale

La **classe modale** est l'intervalle de valeurs qui renferme le plus fort effectif dans une distribution organisée en classes. Autrement dit, c'est la classe qui se manifeste le plus souvent dans un histogramme.

Définition de ggplot2

ggplot est une des librairies du langage R. Elle fait partie du package plotnine, qui facilite la création de graphiques.

Définition de la Population

En statistique, la population se réfère à l'ensemble de tous les individus, objets ou événements qui sont sujets à une étude.

Définition d'un Échantillon

Un échantillon représente une partie de la population étudiée. On fait appel à lui quand il est nécessaire d'analyser l'ensemble de la population en raison de sa complexité.

Définition d'une Variable

Une variable est un attribut quantifiable qui peut varier d'un individu à l'autre.

Définition d'une variable qualitative

Une variable qualitative représente une caractéristique ou une catégorie qui ne peut pas être quantifiée.

Définition d'une variable quantitative

Une Variable Quantitative représente une évaluation numérique et peut être l'objet d'opérations mathématiques.

Définition d'un Intervalle de Confiance

Un intervalle de confiance c'est l'outil qui nous permet d'estimer une plage dans laquelle une valeur inconnue comme une moyenne ou une proportion se situe avec un certain niveau de confiance.

5.2 Description des donnees

Avant de commencer l'étape d'analyse de nos données et leur présentation sous forme de graphiques, nous allons identifier les types de données fournies par notre équipe pédagogique, afin de mener à bien ce projet.

Donc pour cette UE on a eu 3 fichiers CSV contenant des donnees importants ou dans ce rapport on va se baser pour poser notre analyse.

Le fichier incendies renferme des attributs tels que le nom de la commune, le code de la commune, l'année de l'incendie, le mois de l'incendie, la date et l'heure à laquelle l'incendie a été signalé ainsi que les causes principales et secondaires de l'incendie. Ces informations nous seront utiles pour examiner nos données et caractériser ces phénomènes en fonction des problématiques que nous mettrons en place bientôt dans ce rapport.

Le fichier géographique renferme la latitude, la longitude et l'altitude. Ces informations peuvent nous donner l'emplacement précis du lieu de l'incendie, ce qui nous permettra de procéder à des analyses en posant nos questions.

Le dossier Météorologique nous fournira des caractéristiques liées aux conditions météorologiques, ce qui nous permettra de distinguer les divers types de météo au moment de l'incendie.

5.3 Analyse des doneees

Dans cette partie, une fois que toutes les définitions nécessaires sont établies et que notre base de données est complète, nous serons prêts à analyser les données grâce au langage R. Nous emploierons différents types d'histogrammes pour diversifier nos analyses et nous procéderons à une étude approfondie de chaque histogramme.

Nous avons divisé nos problématiques en différentes sections :

5.3.1 Évolution des incendies

5.3.1.1 Évolution des incendies au fil des années Dans cette problématique, on va interroger sur le nombre d'incendies qui se produisent chaque année sur le territoire français. Pour répondre à cette question, il est primordial de définir d'abord certains concepts naturels qui faciliteront l'analyse de cette problématique. Une année se compose de 12 mois et comporte 365 jours. Dans cette étude, nous examinerons l'évolution annuelle du nombre d'incendies à travers tous les départements français, déterminant s'il est en déclin ou en ascension.

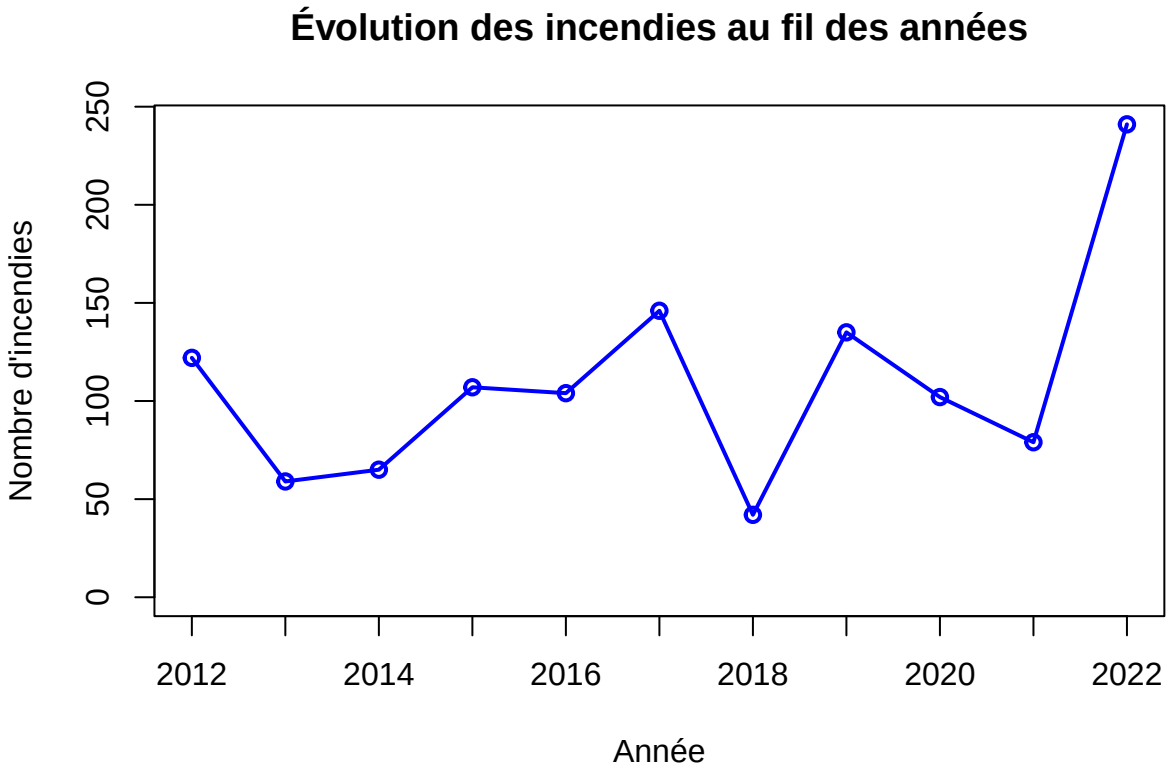
Pour réaliser cette analyse statistique, il faut utiliser la table des **incendies**.Pour réaliser l'histogramme, nous devons utiliser le langage R. Nous avons précisé le chemin d'accès au fichier CSV grâce à une variable que nous avons définie, une variable donnée qui va recevoir la fonction read.csv et le chemin du fichier. Cela permettra à la variable d'accéder au fichier CSV. De plus, pour vérifier notre travail, nous utiliserons la méthode head qui nous donnera les six premières entrées de notre fichier CSV afin de nous assurer que nous travaillons sur le bon fichier et de vérifier également l'en-tête avec les attributs à utiliser.

Pour réaliser cette analyse, nous avons adopté cette technique qui consiste à déterminer la fréquence des incendies par année. En d'autres termes, nous comptabilisons le nombre d'incendies pour chaque année dans la colonne « année » du jeu de données. Cela va nous permettre de créer un tableau qui associe chaque année à son nombre d'incendies.

De plus, nous avons élaboré ce genre de graphique en utilisant un graphique linéaire. Nous avons exploité la fonction « plot() » avec les fréquences obtenues. L'option « o » a été utilisée pour indiquer que nous allions visualiser l'évolution des incendies à la fois par des points et des lignes sur la période donnée.

Pour adapter notre graphique à nos besoins, nous avons opté pour la couleur bleue afin de le rendre plus personnalisable et lisible. De plus, nous avons défini les titres des axes du graphique en utilisant les paramètres xlab, ylab et main.

```r
donnees <- read.csv("../Data/donnees_incendies.csv")

annee_freq <- table(donnees$annee)

plot(annee_freq,
     type="o",  # "o" pour un graphique avec des points et des lignes
     col="blue",
     main="Évolution des incendies au fil des années",
     xlab="Année",
     ylab="Nombre d'incendies")
```



1. **Analyse Informatique:**

Nous allons détailler le processus de développement du code pour structurer notre graphique. Pour cela, nous avons importé les données du fichier grâce à la méthode **read.csv()**. Dans cette méthode, nous avons spécifié le chemin relatif du fichier csv.Par la suite, nous avons déterminé la fréquence des incendies par an en optant pour la colonne **donneees$annee**.En employant la méthode **table()**, nous dénombrons le nombre d'incendies survenant par année.

Cela nous autorisera à constituer un tableau où les indices représentent les années et les valeurs correspondent aux occurrences des incendies.

À présent, nous passons à la création du graphique à l'aide de la méthode **plot()**. Nous avons indiqué la variable **annee_freq** et dans cette méthode, nous avons précisé le type 'o' pour indiquer que les points doivent être affichés et reliés par une ligne. De plus, nous avons défini l'option couleur afin de spécifier que le graphique doit être en bleu.

Avec l'option **main**, nous désignons le titre du diagramme. Ensuite, grâce à xlab et ylab, nous définissons les étiquettes des axes x et y, où x représente l'**année** et y représente le **nombre d'incendies**.

2. **Analyse Statistique:**

Alors, débutons par cette étude ; nous allons mener notre analyse sur une période de 11 ans, de 2012 à 2022.

Nous allons donc commencer à dresser le bilan de nos effectifs d'incendies. En 2012, nous avons enregistré 122 incidents, en 2013, 59 incendies, en 2014, 65 incendies, en 2015, le nombre a grimpé à 107. En 2016, nous avons eu 104 incendies. Pour l'année suivante, en 2017, nous avons connu une augmentation avec 146 incidents. Puis en 2018, le chiffre est redescendu à 42 incendies. En 2019, nous avons rapporté 135

incendies et pour l'année suivante, en 2020 nous avions enregistré 102 incidents. Enfin pour2021 ,nous avons comptabilisé 79 incendies et pour2022 ,le nombre d'incendies a fortement augmenté atteignant 241incendies .

Il est possible d'observer que la moyenne générale des incendies en France sur une durée de 11 ans s'élève à 109,27. Maintenant que nous avons compilé les totaux des onze années consécutives, nous allons détailler notre analyse. En démarrant de 2012, nous constatons un total de 122 incendies, ce qui pourrait être considéré comme « bon ». Cependant, cette comparaison est prématurée tant que nous n'avons pas répertorié les autres chiffres pour une évaluation plus approfondie. Après une baisse de 63 incendies en 2013, on a constaté une augmentation de 6 incendies en 2014 lors de la comparaison entre 2013 et 2014.

Cependant, en 2015, le nombre d'incendies a considérablement augmenté, atteignant 107 incendies pour l'année 2015, ce qui représente un chiffre élevé mais pas le plus élevé que nous ayons connu. En comparaison avec 2012, nous avons atteint un taux qui est le deuxième record maximal. Puis en 2016, ce taux a baissé de trois incendies par rapport à l'année précédente.

En 2017, on a observé une hausse remarquable du taux d'incendies qui a atteint un nouveau sommet à 146, soit une augmentation de 42 incendies par rapport à 2016. Si l'on compare avec l'année 2012, cela représente une différence de 24 incendries. En 2018, on constate une baisse rapide du nombre d'incendies, enregistrant un taux de 42 qui, selon les données de 2018, représente le plus bas enregistré depuis 2012. En 2019, on a noté une hausse significative du nombre d'incendies, atteignant 135, soit une différence de 93 incendies par rapport à 2018.

En 2020, une baisse du nombre d'incendies entraîne une variation de 33 incendies par rapport à l'année 2029. Ensuite, on observe une baisse en 2021 avec une différence de 23 incendies par rapport à l'année 2020. Finalement, un chiffre très élevé en 2022 a été atteint, atteignant le nombre record de 241 incendies. De plus, cela représente une différence de 162 incendres par rapport à l'année 2021.

On peut résumer qu'au cours de ces 11 années, l'année où le nombre d'incendies est le plus élevé est 2022 avec un total de 241 incendies, tandis que l'année où ce nombre est le plus bas est 2018.

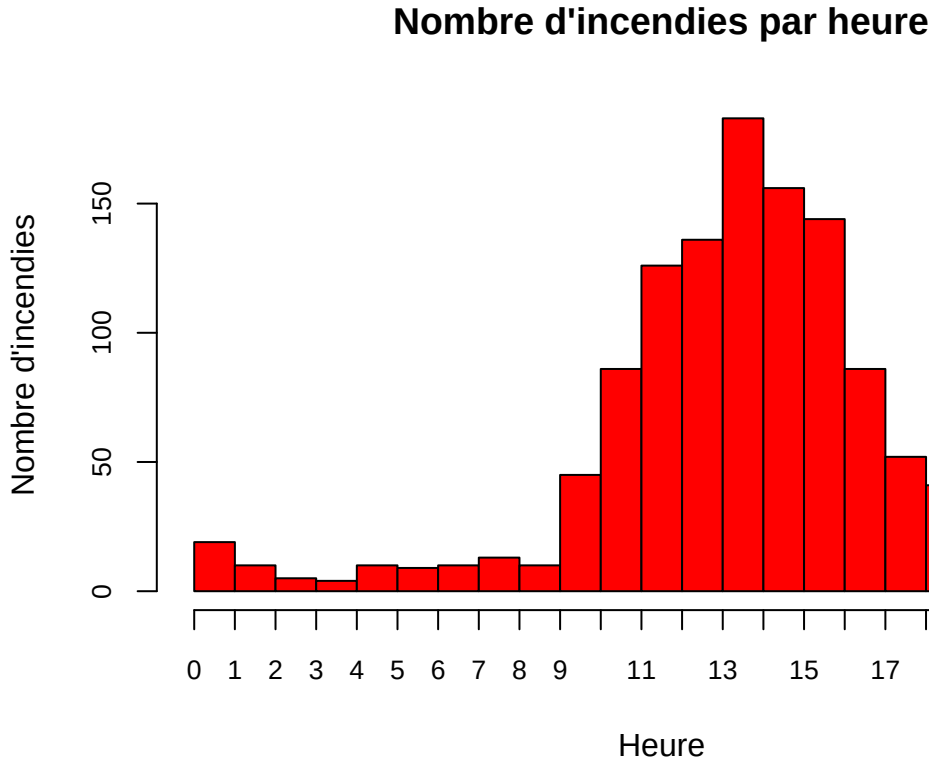
En 2022, on a observé une augmentation atypique, avec un total de 241 incendies. Divers éléments peuvent expliquer cette augmentation, tels que les modifications extrêmes des conditions météorologiques. En 2022, des vagues de chaleur sévères et une sécheresse persistante ont contribué à cet accroissement. De plus, les actions humaines, notamment la négligence, ont influencé cette hausse.

Pour conclure notre analyse, l'évolution des incendies de 2012 à 2022 a montré des tendances significatives et atypiques. Cela souligne l'importance d'examiner l'impact des facteurs climatiques, humains et socio-économiques sur le nombre d'incendies. On peut conclure que l'année 2022 a été exceptionnelle, affichant le plus haut nombre d'incendies. De plus, nous avons observé une variabilité des tendances, comme démontré souvent avec les années 2017, 2014 et 2013 où les taux étaient remarquablement bas comparés à ceux de 2022. Il est essentiel de noter que le changement climatique a une grande importance.

Enfin, et c'est crucial, la gestion et la prévention jouent un rôle primordial. Les préfectures, ainsi que les forces de police municipales et nationales, peuvent contribuer à maîtriser ce taux d'incendies attribués à des actes de malveillance ou de négligence.

```r
incendies <- read.csv("../Exports/export_incendies.csv")
incendies$heure <- as.character(incendies$heure)
incendies$heure <- as.integer(sub(":.*", "", incendies$heure))
hist(
  incendies$heure,
  main = "Nombre d'incendies par heure",
  xlab = "Heure",
  ylab = "Nombre d'incendies",
```

```
  col = "red",
  border = "black",
  breaks = seq(0, 23, 1),
  xaxt = "n",
  cex.axis = 0.8
)
axis(1, at = seq(0, 23, 1), labels = seq(0, 23, 1), cex.axis = 0.8)
```



5.3.1.2 Analyse des heures d'incendie

1- **Analyse Informatique**:

Le code débute en chargeant les données à partir du fichier CSV via l'instruction **incendies <- read.csv("donnees_incendies.csv")**.

Cette étape est indispensable pour importer les données relatives aux incendies et les organisent sous forme de tableau, un format utilisé par R pour gérer les informations structurées.

Par la suite, nous **convertissons** la colonne de l'heure des incendies : **incendies**$heure = as.character(incendies$**heure)**

La **colonne** doit être convertie en **chaîne de caractères**, car certaines opérations de traitement textuel ne peuvent se faire qu'avec ce format de données.

L'instruction suivante, **incendies**$heure = as.integer(sub(" : .","",incendies$**heure))**, utilise la fonction **sub(":", " ", incendies$heure)** pour éliminer tout ce qui suit les deux-points « : », ne conservant ainsi que l'heure entière.

La fonction **as.integer()** transforme par la suite cette valeur en entier, facilitant ainsi son emploi dans **des calculs statistiques et l'élaboration de l'histogramme.**

Par la suite, la fonction **hist()** est employée pour **créer l'histogramme**, en utilisant **hist(incendies$heure,...)**.

L'argument **main="Nombre d'incendies par heure"** attribue un titre au diagramme, alors que **xlab="Heure*" **et** ylab="Nombre d'incendies"** déterminent les désignations des axes pour spécifier les données présentées.

L'option **col="red"** permet de remplir les barres avec une teinte rouge pour une distinction colorimétrique plus efficace tandis que **border="black"** génère des bordures noires pour délimiter visuellement les barres.

L'option **breaks=seq(0, 23, 1)** segmente l'axe des heures en intervalles d'une unité par barre dans le graphique, chaque barre symbolisant une heure.

Pour finir, la commande **axis(1,at=seq(0,23,1),labels=seq(0,23,1))** permet de personnaliser l'axe des x. Initialement, l'option **xaxt="n"** dans **hist()** avait éliminé l'affichage automatique des étiquettes sur cet axe.

Cette ligne finale ajoute **manuellement les vingt-quatre heures** afin d'améliorer l'harmonie esthétique.

Cela garantit une clarté accrue du graphique et facilite une interprétation optimale des données.

2- **Analyse Statistique** :

Ce graphique montre une analyse du nombre d'incendies à des heures précises sur le **territoire français.**

On analyse la fréquence des feux sur une durée de **24 heures.**

L'axe X illustre l'ensemble des heures, de **0 à 23**, tandis que l'axe Y représente le total des incendies, avec une portée d'environ **0 à près de 200.**

Nous avons développé **un histogramme** comme type de graphique.

On observe un pic significatif **entre 11 heures et 14 heures**, atteignant son maximum aux alentours de **12-13** heures avec environ **200 incendies**.
Après le pic du midi, l'occurrence des incendies commence à décroître graduellement jusqu'à **23 heures**, avec quelques variations mineures, comme un léger rebound autour de **18-19 heures.**

Cette répartition des données est asymétrique, montrant une queue à droite après le pic principal de **12-13 h**, car la **fréquence** diminue plus progressivement vers les heures tardives **14-23 h** qu'elle ne s'accroît avant le pic.

Cette répartition indique une connexion forte entre les **actions humaines** et les incendies.

En effet, les moments où les incendies sont le plus courants correspondent à des périodes **d'activité humaine intense**, particulièrement entre **midi et 16h.**

Ceci pourrait être attribué à une utilisation croissante d'équipements électriques, de cuisines ou encore de travaux dans l'industrie ou l'agriculture, augmentant ainsi les dangers d'incendie.

L'explication de la **rareté des incendies** durant la nuit et le matin réside dans la diminution des interactions humaines et l'utilisation limitée de sources de chaleur et d'énergie.

Par contre, l'augmentation en pleine journée pourrait résulter des températures estivales chaudes et de l'intensification des activités humaines durant cette tranche horaire.

Par après 16 heures, le nombre d'incendies commence à baisser peu à peu, ce qui peut être attribué à un ralentissement des occupations à risque et à une prise de conscience accrue des dangers suite au pic observé.

Cependant, on note un léger rebond vers **20 heures**, probablement associé aux tâches de fin de journée telles que la cuisine ou l'utilisation d'appareils électriques avant le coucher.

En définitive, cette répartition horaire des incendies met en exergue l'influence des comportements humains sur leur fréquence.

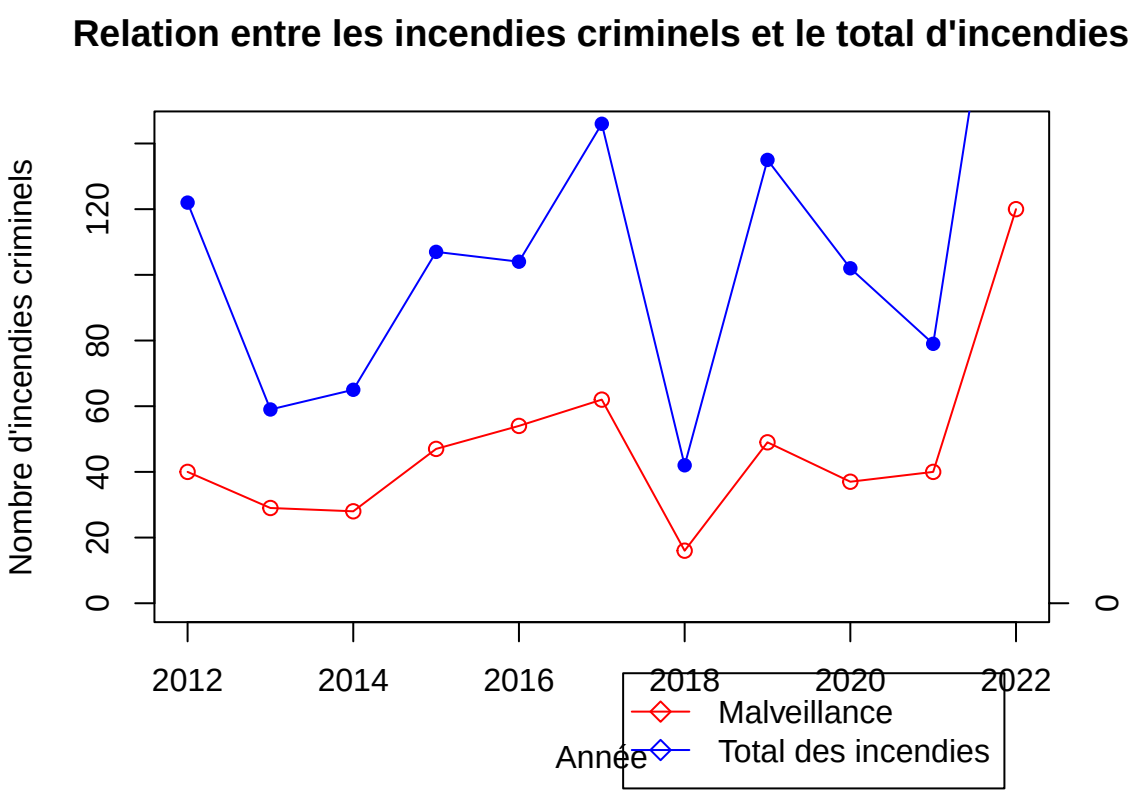
Ces données pourraient s'avérer utiles pour optimiser la prévention et l'intervention des services d'urgence en ciblant les périodes critiques de la journée.

Le **mode** principale se situe autour de 12-123 heures, moment où le nombre d'incendies atteint approximativement 200.

Il est probable que **la médiane**, c'est-à-dire le moment où **50% des incendies** se produisent, se situe aux alentours de **11-12 heures**, car la majorité d'entre eux ont lieu **après 14h.**

5.3.1.3 Évolution décennale des causes d'incendies Dans cette problematique on va analyser l'evolution decennale des causes d'incendies

```
incendies <- read.csv("../Data/donnees_incendies.csv")

incendies$annee <- as.numeric(incendies$annee)


incendies_criminels <- subset(incendies, nature_inc_prim == "Malveillance")

incendies_par_annee <- table(incendies_criminels$annee)

incendies_par_annee_df <- data.frame(annee = as.numeric(names(incendies_par_annee)),
                                     nombre_incendies = as.vector(incendies_par_annee))


incendies_total_par_annee <- table(incendies$annee)

incendies_total_par_annee_df <- data.frame(annee = as.numeric(names(incendies_total_par_annee)),
                                           nombre_incendies_total = as.vector(incendies_total_par_anne

  par(mar=c(5, 4, 4, 5) + 0.1)

  plot(incendies_par_annee_df$annee, incendies_par_annee_df$nombre_incendies,
       type="o", col="red",
       xlab="Année", ylab="Nombre d'incendies criminels",
       main="Relation entre les incendies criminels et le total d'incendies",
       ylim=c(0, max(incendies_par_annee_df$nombre_incendies) * 1.2))  # Aumenta el límite del eje y iz

  lines(incendies_total_par_annee_df$annee, incendies_total_par_annee_df$nombre_incendies_total,
        type="o", col="blue", pch=16)

  axis(4, at=seq(0, max(incendies_total_par_annee_df$nombre_incendies_total), by=500),
       labels=seq(0, max(incendies_total_par_annee_df$nombre_incendies_total), by=500))

  legend("topright", legend=c("Malveillance", "Total des incendies"),
         col=c("red", "blue"), lty=1, pch=5, xpd=TRUE, inset=c(0.05, 1.1))
```



Analyse Informatique:

La variable “année” a été modifiée en un format numérique afin de permettre son exploitation sans erreur dans les analyses à venir :

incendies*annee* $<-$ *as.numeric*(*incendies***annee**)

Les incendies dont la cause principale est répertoriée comme étant la “malveillance” ont pu être isolés grâce à la commande :

incendies_criminels <- subset(incendies, nature_inc_prim == “Malveillance”)

Cela permet d’examiner plus en détail cette catégorie d’incendies.

La fonction table() a facilité le comptage des incendies criminels sur une base annuelle. Ces résultats ont par la suite été convertis en dataframe pour en faciliter une exploitation graphique ultérieure :

incendies_par_annee_df <- data.frame(annee = as.numeric(names(incendies_par_annee)), nombre_incendies = as.vector(incendies_par_annee))

Une tâche similaire a été réalisé pour tous les types d’incendies sans exception :

incendies_total_par_annee_df <- data.frame(annee = as.numeric(names(incendies_total_par_annee)), nombre_incendies_total = as.vector(incendies_total_par_annee))

Cela facilitera la comparaison avec uniquement les incendies criminels.

Il sera désormais possible de produire des graphiques pour une comparaison visuelle entre le nombre total d’incendies et celui des incendies criminels.

Pour illustrer la comparaison entre les incendies criminels et le total des incendies annuels, nous élaborons un diagramme où les points symbolisant les incendies criminels sont reliés par une ligne de couleur rouge.

Tous les points sont pris en compte par l'ajustement de l'axe Y.

Par la suite, nous incluons la courbe bleue représentant le total des incendies, avec des points plus distincts.

Cela facilite l'identification des corrélations entre les pics d'incendies criminels et les pics totaux.

Étant donné la difficulté de comparer les grandes valeurs du total sur un seul axe Y, nous ajoutons un second axe à droite pour cette courbe.

Pour finir, nous positionnons la légende à la droite, en dehors du diagramme, en désignant la ligne rouge comme représentant les incendies criminels et la bleue comme le total.

Cette mesure rend les données présentées plus claires tout en préservant une excellente lisibilité.

Ainsi, le graphique offre la possibilité d'analyser la corrélation entre les incendies criminels et la progression générale des incendies au cours des années.

Analyse Statistique:

Nous avons analysé le lien entre les incendies criminels intentionnels et le total des incendies sur la période de 2012 à 2022.

L'axe des abscisses représente les différentes années tandis que l'axe des ordonnées indique le nombre de feux.

On observe deux courbes :
- En **rouge**, l'évolution variable du nombre d'incendies intentionnels (causés par la malveillance humaine).

- En **bleu**, le total variable des incendies relevés chaque année.

Fluctuation du nombre total de feux :

On observe une variation significative du total des incendies, atteignant des pics en 2018 et 2022.

Ces augmentations pourraient être dues à des conditions climatiques extrêmes ou à des phases prolongées de sécheresse qui favorisent les incendies.

Évolution des incendies criminels :

Le nombre d'incendies criminels présente une tendance semblable, mais avec une portée réduite.
Nous constatons une hausse marquée en 2022, qui pourrait signaler une intensification des actes de malveillance intentionnelle.

Corrélation partielle observée entre les deux graphiques :

Bien que les incendies criminels suivent globalement la tendance générale du nombre total de feux, quelques déviations demeurent.

Par exemple, en 2018, on observe une augmentation notable du total des incendies, alors que la progression des feux d'origine criminelle reste plus mesurée.

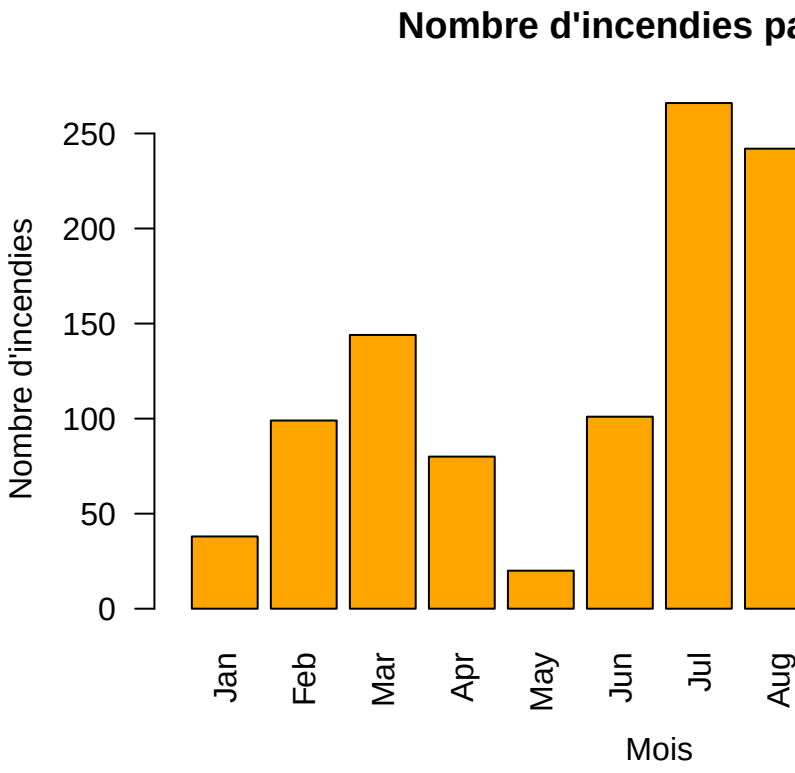
Ceci pourrait indiquer que d'autres éléments (météo, accidents) ont eu un impact significatif sur l'augmentation générale des incendies cette année-là.

L'étude indique que les incendies intentionnels constituent une part importante de l'ensemble des feux, mais leur progression ne coïncide pas toujours parfaitement avec la même tendance.

Les fluctuations des incendies pourraient être affectées par certains phénomènes climatiques ou des phases de sécheresse prolongées, alors que les feux allumés volontairement sont plus en lien avec des éléments sociétaux et de sécurité.

Ces informations sont indispensables pour guider les stratégies de prévention des incendies, en distinguant les actions à entreprendre en fonction de la provenance des flammes.

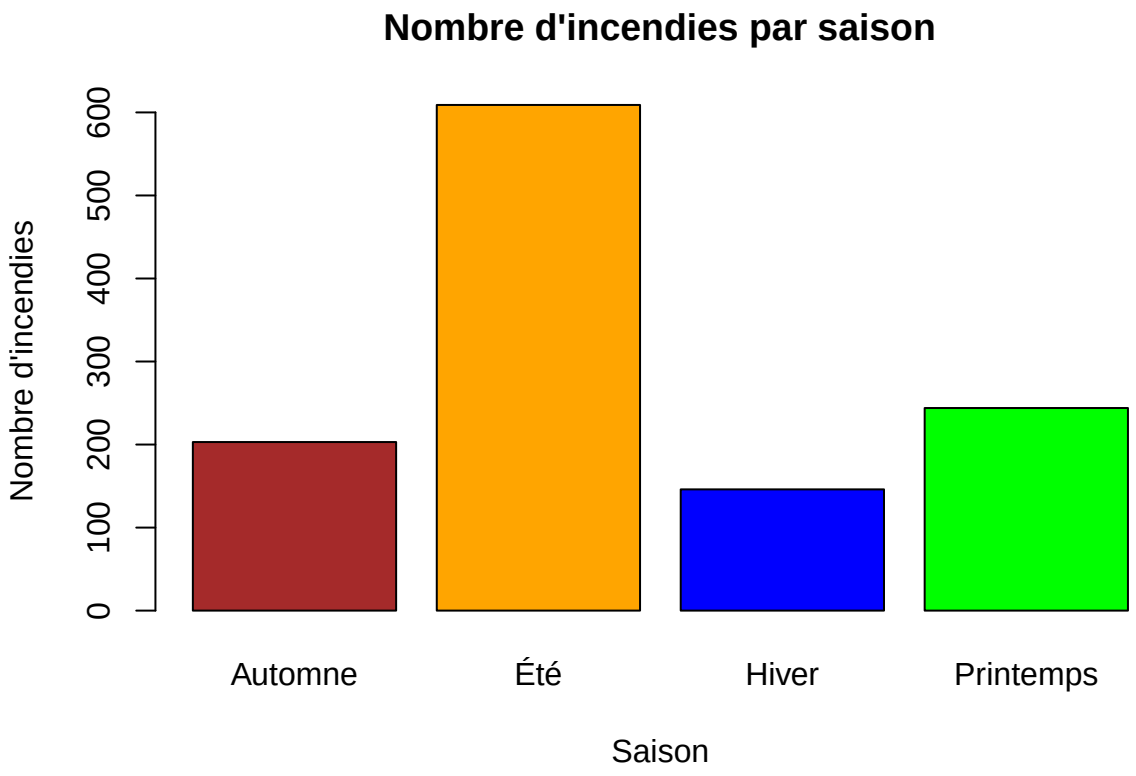
```r
incendies <- read.csv("../Data/donnees_incendies.csv")
incendies$mois <- factor(incendies$mois, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun",
                                                    "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))

nb_incendies_par_mois <- table(incendies$mois)

barplot(nb_incendies_par_mois,
        col = "orange",
        border = "black",
        main = "Nombre d'incendies par mois",
        xlab = "Mois",
        ylab = "Nombre d'incendies",
        las = 2)
```



5.3.1.4 Analyse des incendies par mois et saison

```r
incendies$saison <- NA

incendies$saison[incendies$mois %in% c("Dec", "Jan", "Feb")] <- "Hiver"
incendies$saison[incendies$mois %in% c("Mar", "Apr", "May")] <- "Printemps"
incendies$saison[incendies$mois %in% c("Jun", "Jul", "Aug")] <- "Été"
incendies$saison[incendies$mois %in% c("Sep", "Oct", "Nov")] <- "Automne"

nb_incendies_par_saison <- table(incendies$saison)

couleurs_saisons <- c("Hiver" = "blue", "Printemps" = "green", "Été" = "orange", "Automne" = "brown")
```

```r
barplot(nb_incendies_par_saison,
        col = couleurs_saisons[names(nb_incendies_par_saison)],
        border = "black",
        main = "Nombre d'incendies par saison",
        xlab = "Saison",
        ylab = "Nombre d'incendies")
```



1- **Analyse Informatique:**

Pour débuter, nous arrangeons les mois selon l'ordre désiré :
incendies$mois <- factor(incendies$**mois, levels = c("Jan", "Fév", "Mars", "Avr", "Mai", "Juin", "Juil", "Août", "Sept", "Oct", "Nov", "Décembre")).**

Cette phase convertit la colonne mois en un facteur catégorique, établissant clairement la séquence des mois de janvier à décembre.

Ceci assure un classement précis des mois lors de la création de graphiques ou de calculs.

Ensuite, nous calculons le nombre d'incendies par mois avec la commande **nb_incendies_par_mois <- table(incendies$mois)**.

Cette opération crée un tableau de décompte des incendies par mois.

Un diagramme à barres est donc généré pour représenter le nombre d'incendies mensuels en utilisant **barplot(nb_incendies_par_mois,...)**.

Dans cette fonction :

- **col="orange"** : Les barres sont colorées en orange.

- **border="noir"** : Ajoute une bordure noire autour des barres.
- **main="Nombre** d'incendies par mois" : Le titre du graphique.
- **xlab="Mois"** et **ylab="Nombre d'incendies"** : Les étiquettes des axes.
- **las=2* : Cette option fait pivoter les étiquettes des mois pour une meilleure lisibilité, surtout si les noms sont longs.

Par la suite, une colonne saison est ajoutée à l'objet incendie pour assigner une saison à chaque mois.

L'instruction **incendies$saison <- NA** initialise d'abord une colonne vide, puis les lignes subséquentes attribuent une saison à chaque mois en se basant sur les valeurs de la colonne mois :

- Déc, Jan, Fév : **Hiver**
- Mars, Avr, Mai : **Printemps**
- Juin, Juil, Août : **Été**
- Sept, Oct, Nov : **Automne**

Après avoir analysé les données sur les incendies, nous avons calculé le nombre d'incendies qui se sont produits pendant chaque saison à l'aide de la fonction **table()**.

Cela nous a permis de générer un tableau croisé indiquant la fréquence des incendies selon la période de l'année à laquelle ils se sont déclarés.

2- **Analyse Statistique:**

Nous avons examiné le total d'incendies sur une durée de douze mois, de janvier à décembre. L'axe horizontal indique les différents mois de l'année, alors que l'axe vertical dépeint le nombre total d'incendies, qui varie selon les périodes.

Le graphique illustre donc **les variations mensuelles**.

Pour visualiser facilement les variations, nous avons choisi **un diagramme en barres**.

De **janvier à mars**, on observe une montée progressive du nombre de feux, vraisemblablement due aux conditions hivernales qui accroissent les dangers liés à la chaleur et l'usage des dispositifs de chauffage.

Cette évolution continue en **février et mars**.

En **mai**, le nombre d'incendies est relativement bas, probablement à cause d'un temps plus clément et d'une diminution des activités potentiellement dangereuses avant l'arrivée de l'été.

De **juin à août**, on observe une montée, surtout en juillet et août, qui sont les mois les plus chauds où les activités humaines et agricoles se multiplient, provoquant de nombreux incendies.

De **septembre à décembre**, les risques d'incendie diminuent progressivement avec l'arrivée des mois plus frais et humides, réduisant ainsi les dangers.

Cette répartition met en évidence l'influence du climat et des actes humains en fonction des saisons, entre les besoins en chauffage d'hiver et estivaux, augmentant ainsi les risques, à l'opposé de la période de fin d'année.

Le **second graphique** propose une étude de la **récurrence des feux en fonction des saisons sur le sol français.**

Les données sont distribuées selon les **quatre saisons classiques** : l'hiver, le printemps, l'été et l'automne.

L'axe horizontal illustre ces diverses saisons alors que l'axe vertical indique le total d'incendies pour chaque saison.

Pour représenter la distribution saisonnière des incendies, nous avons choisi d'utiliser un diagramme à barres.

L'été se distingue de manière significative avec un nombre d'incendies pratiquement triplé par rapport aux autres saisons.

Cela est dû aux températures élevées de l'été et à la hausse des dangers d'incendies de forêt. Cette augmentation est également alimentée par l'intensification des activités humaines pendant l'été, qu'il s'agisse de la cultivation ou des divertissements en plein air.

Le nombre d'incendies au printemps est généralement inférieur mais plus élevé qu'en automne et en hiver.

Ceci pourrait être dû à la fin de la saison hivernale combinée aux premiers pics de chaleur, entraînant quelques incendies, en particulier d'origine agricole.

Même si le nombre d'incendies à l'automne est moins élevé qu'en été, il reste néanmoins important.

Des températures plus clémentes diminuent le danger des méga-incendies, toutefois les travaux agricoles et la chute des feuilles continuent d'être des éléments à risque.

L'hiver est marqué par une incidence réduite d'incendies grâce à des températures basses et un taux d'humidité plus élevé qui minimisent les dangers.

L'usage limité de certains appareils, tels que les cheminées, ainsi que la réduction des activités en plein air contribuent aussi à ce nombre réduit.

En définitive, l'analyse saisonnière des incendies met en évidence l'impact du climat et des actions humaines sur leur occurrence.

L'été est la saison la plus risquée, tandis que l'hiver enregistre les statistiques les plus faibles.

Ces informations sont indispensables aux services de prévention et d'intervention pour identifier les saisons à haut risque et mettre en place des dispositifs de sécurité appropriés.

Ainsi, le maximum de fréquence se produit durant l'été.

Il est probable que la médiane, c'est-à-dire la saison durant laquelle 50% des incendies se produisent, se situe entre le printemps et l'été étant donné la forte concentration estivale.

5.3.1.5 Heures critiques de déclenchement

```
incendies <- read.csv("../Data/donnees_incendies.csv")

incendies$jour <- as.Date(incendies$jour, format="%Y-%m-%d")

head(incendies$jour)
```

5.3.1.6 Cyclicité hebdomadaire

```
## [1] "1970-01-29" "1970-01-30" "1970-01-03" "1970-01-03" "1970-01-09"
## [6] "1970-01-15"
```

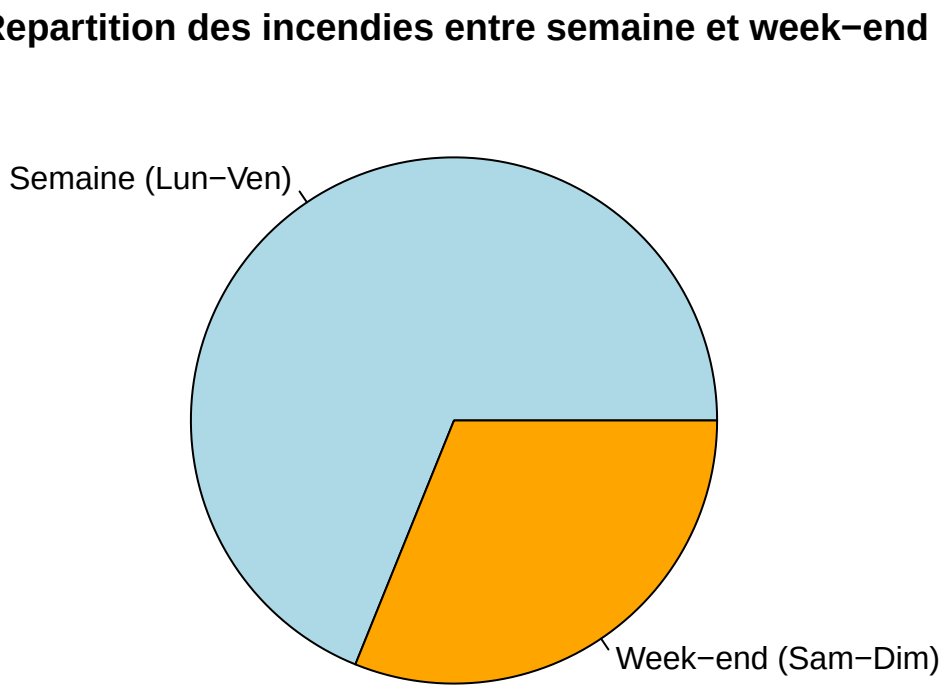
```
Sys.setlocale("LC_TIME", "fr_FR.UTF-8")
```

```
## [1] "fr_FR.UTF-8"
```

```
incendies$jour_semaine <- weekdays(incendies$jour, abbreviate = FALSE)

print(unique(incendies$jour_semaine))
```

```
## [1] "jeudi"    "vendredi" "samedi"   "dimanche" "mercredi" "mardi"    "lundi"
```

```r
incendies$jour_semaine <- factor(incendies$jour_semaine,
                                 levels = c("lundi", "mardi", "mercredi", "jeudi", "vendredi", "samedi"

semaine_incendies <- sum(incendies$jour_semaine %in% c("lundi", "mardi", "mercredi", "jeudi", "vendredi
weekend_incendies <- sum(incendies$jour_semaine %in% c("samedi", "dimanche"))

total_incendies <- c(semaine = semaine_incendies, weekend = weekend_incendies)

par(mar = c(2, 2, 2, 2))
pie(total_incendies,
    col = c("lightblue", "orange"),
    main = "Repartition des incendies entre semaine et week-end",
    labels = c("Semaine (Lun-Ven)", "Week-end (Sam-Dim)"),
    cex = 1)
```



1. **Analyse Informatique:**

Transformation de la colonne "date" en type Date:

Nous utilisons la fonction as.Date() pour transformer la colonne « date » en un objet de type Date. Le format indiqué est **"%Y-%m-%d"**, qui représente **l'année, le mois et le jour selon la norme**.

Extraction du jour de la semaine:

La fonction **weekdays()** est employée pour extraire **le nom du jour de la semaine** à partir de la colonne **date**.

Nous commençons par définir la langue française de localisation grâce à **Sys.setlocale()**, afin que les jours soient affichés dans cette langue, comme par exemple « lundi ».

Verification des jours uniques:

Cette instruction permet de montrer les valeurs distinctes trouvées dans la colonne « jour_semaine » pour vérifier les jours récupérés.

Création d'un facteur pour les jours:

Nous convertissons la colonne « jour_semaine » en facteur, avec les niveaux des jours clairement définis de lundi à dimanche, assurant ainsi leur séquence correcte.

Calcul du nombre d'incendies pendant la semaine et le weekend:

Nous comptons ici les incendies qui se déclenchent durant **la semaine, du lundi au vendredi**, ainsi que pendant **le week-end**, c'est-à-dire **samedi et dimanche**.

Ceci est effectué en utilisant l'opérateur **%in%** pour confirmer quel jour correspond à l'événement.

Création d'un vecteur avec les résultats:

Un tableau nommé **« total_incendies »** est créé pour conserver **les résultats** des incendies durant la **semaine et le week-end**.

Creation d'un graphique circulaire:

Pour finir, nous utilisons la fonction **pie()** pour créer un **diagramme circulaire** afin d'illustrer la répartition des incendies entre la semaine et le weekend, en optant pour des couleurs et étiquettes sur mesure, tout en modifiant les marges pour une présentation plus soignée.

2. **Analyse Statistique:**

Incendies fréquents pendant la semaine active:

Les informations indiquent que la majorité des incendies se déclarent en semaine, à cause de divers facteurs.

Parfois, l'augmentation de l'activité humaine et une forte densité de population peuvent engendrer plus de risques.

Au cours des heures de travail et durant les déplacements, les activités domestiques ou industrielles augmentent les risques potentiels.

De plus, les incendies peuvent également être favorisés par le travail ou les conditions météorologiques : l'usage d'instruments ou d'équipements électriques lors de journées de travail intenses en est fréquemment une raison.

Fin de semaine plus paisible:

Le week-end semble présenter un risque d'incendie réduit, en raison de la diminution des activités professionnelles et industrielles, ainsi que de la réduction des déplacements et de l'utilisation d'appareils susceptibles de déclencher un feu.

Il est probable qu'en week-end, lorsque les activités commerciales et industrielles sont réduites ou absentes, les personnes utilisent moins d'appareils ou adoptent davantage de mesures de précaution.

5.3.2 Facteurs climatiques et météorologiques

5.3.2.1 Influence de la température sur les incendies

5.3.2.2 Impact de l'humidité sur les incendies Pour mener l'analyse de l'influence de l'humidité sur les incendies, il est nécessaire de constituer notre base d'analyse. Il est nécessaire d'utiliser les deux tables que nous avons mises en place dans notre Base de données, à savoir la Table Incendies et la Table donnees_meteo.

Afin de réaliser une jointure entre ces deux tables, nous avons fait appel à une troisième table nommée humidite qui regroupe les champs des deux tables visées, partageant un élément en commun : le « Code_INSEE ». Nous avons détaillé la méthode utilisée pour cela dans la section Informatique de notre rapport.

Avant d'approfondir dans les détails de notre étude, nous allons d'abord définir les termes clés que nous utiliserons dans notre analyse. L'humidité se réfère à la présence de vapeur d'eau dans l'air.

Le but de notre analyse est d'étudier le lien entre l'humidité atmosphérique et les incendies. Pour accomplir cela, nous devons étudier la relation entre l'humidité et la dimension des incendies.

Il est essentiel de mettre en évidence deux attributs importants.

1. **Tens_vap_med**: Cet attribut évalue la pression de vaporisation moyenne, qui est une autre façon de dire qu'il s'agit d'un indicateur de l'humidité de l'air.
2. **surface_parcourue_m2**: Cette caractéristique présente la superficie ravagée par un feu, ce qui en fait un instrument pour évaluer l'intensité du feu.

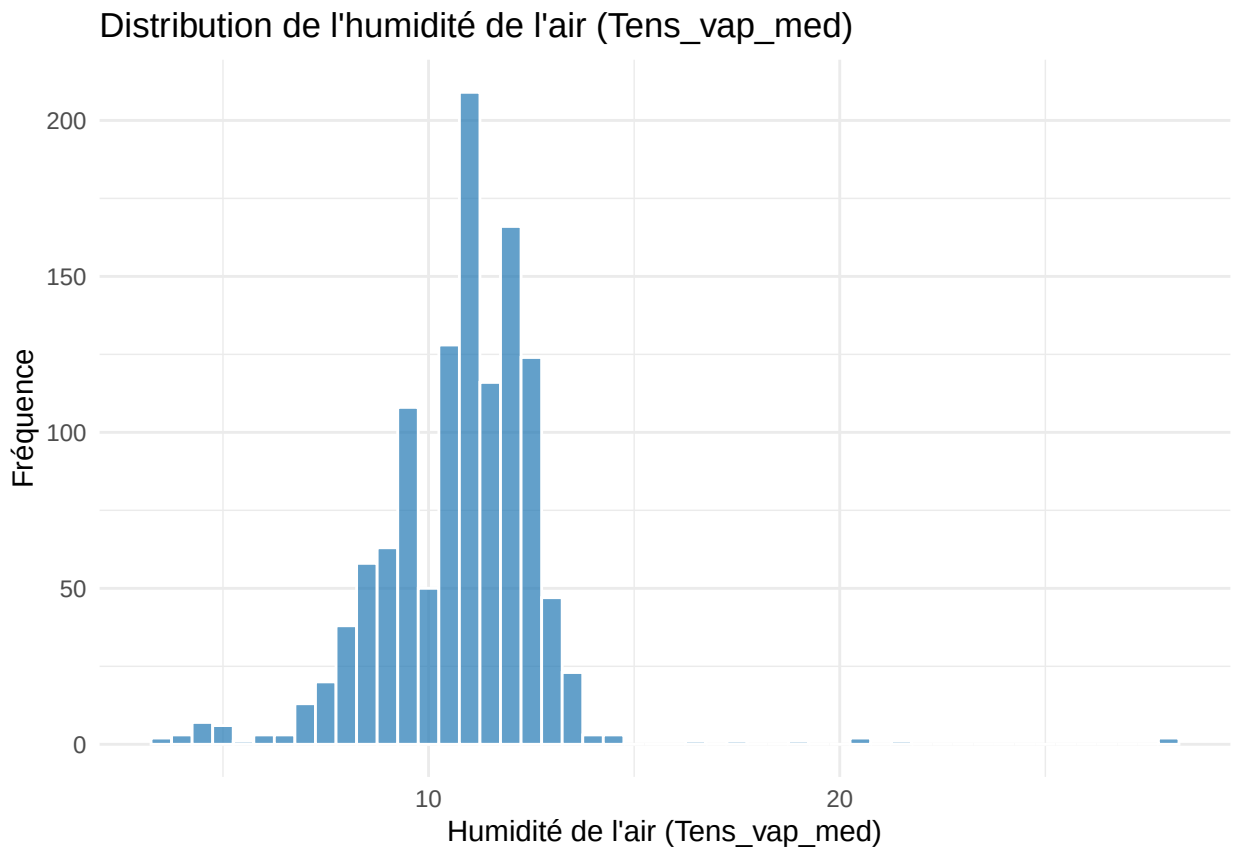
Dans cette étude, nous allons créer et analyser deux graphes indispensables à notre problématique.

1. **Histogramme de l'humidité**

Le diagramme de l'humidité nous aidera à examiner et à saisir la distribution des taux d'humidité dans l'échantillon de données.

L'humidité joue un rôle crucial dans l'analyse des incendies en raison de son impact sur la rapidité de leur avancement. Cependant, avant d'examiner toute relation avec la surface incendiée, nous devons d'abord comprendre comment l'humidité fluctue dans les données.

```
library(ggplot2)
data <- read.csv("../Exports/export_Humidites.csv")
ggplot(data, aes(x = Tens_vap_med)) +
  geom_histogram(binwidth = 0.5, fill = "#1f77b4", color = "white", alpha = 0.7) +
  labs(title = "Distribution de l'humidité de l'air (Tens_vap_med)",
       x = "Humidité de l'air (Tens_vap_med)",
       y = "Fréquence") +
  theme_minimal()
```



1. **Analyse Informatique:**

Dans notre code destiné à la création du graphique, nous avons employé le langage R pour sa réalisation. Tout d'abord, nous avons importé le fichier en utilisant la méthode **read.csv()**.

Par la suite, on procède à l'initialisation du graphique que l'on souhaite concevoir avec la méthode **ggplot()**.On passe en paramètres de cette méthode les données ainsi que l'**aes**,qui nous permet de spécifier que l'axe des x sera représenté par **Tens_vap_med**.

On définit ensuite une autre méthode **geom_histogram()** pour intégrer un histogramme au graphique. Dans cette méthode, on spécifie des paramètres pour déterminer la largeur des barres de l'histogramme, et par la suite, on remplit les barres en bleu. Nous définissons la bordure en blanc et rendons les barres davantage transparentes.

On détermine les titres et étiquettes en faisant appel à la méthode **labs()**, en exploitant les paramètres title, x et y.Le titre indique le sujet du graphique, 'x' correspond à l'étiquette de l'axe horizontal et 'y' correspond à l'étiquette de l'axe vertical.

Et pour donner un aspect minimaliste au thème, nous avons employé la méthode **theme_minimal()** afin d'incorporer un style plus contemporain.

2. **Analyse Statistique:**

L'axe des abscisses illustre l'humidité de l'air, déterminée par la pression de vapeur moyenne. Les chiffres se situent approximativement entre 0 et 25.On présume que les valeurs sont exprimées en hPa. C'est une unité utilisée pour évaluer la pression de la vapeur.

L'axe ordonnes illustre la fréquence, c'est-à-dire le total des observations pour chaque plage de tension de vapeur. Il a atteint une fréquence maximale de 200.

D'après les informations intégrées, cet histogramme révèle une distribution asymétrique avec une importante concentration de valeurs faibles en matière d'humidité de l'air, allant de 5 à 15 hPa.

L'histogramme présente une asymétrie vers la droite, avec un grand nombre d'observations ayant de faibles valeurs de tension de vapeur, indiquant une humidité faible à modérée, et quelques observations à des valeurs plus élevées, traduisant une humidité supérieure.

On pourrait affirmer que la classe modale de cet histogramme se situe approximativement entre 10 et 12 hPa.

La portée de cet histogramme s'étend de 0 à 25 hPa. Néanmoins, on remarque qu'il existe très peu de données entre 20 et 25 hPa. De plus, il est évident qu'au-delà de 22 hPa, aucune observation n'est présente.

Dans notre histogramme, on peut observer la présence d'une longue queue, bien qu'elle soit peu dense. Autrement dit, cela nous indique que les valeurs élevées de tension de vapeur sont peu fréquentes dans cet ensemble de données.

Nous allons déterminer les interprétations des intervalles concernant le taux d'humidité.

1. **0 a 5 hPa**: Cette plage indique un taux d'humidité très bas.
2. **5 a 15 hPa**: C'est la zone où est rassemblée la plupart des données. La fréquence s'accroît rapidement dès que l'on atteint 5 hPa, atteignant un maximum aux alentours de 10 à 12 hPa avant de redescendre. Cela signifie que le niveau d'humidité est modéré.
3. **15 a 20 hPa**:Dans cette période, nous observons une réduction qui est associée à des conditions plus humides.
4. **20 a 25 hPa**:Les observations sont peu fréquentes. La fréquence étant pratiquement de 0, on peut observer que nous avons des conditions hors du commun, telles que des climats tropicaux.

Dans un cadre météorologique, la tension de vapeur représente une évaluation de la pression partielle de la vapeur d'eau dans l'atmosphère, liée directement à l'humidité. Selon notre histogramme, une pression de vapeur de 10 hPa est associée à une humidité relative modérée, tandis qu'une pression de vapeur avoisinant les 20 hPa indique un niveau d'humidité considérablement plus élevé, potentiellement proche de la saturation.

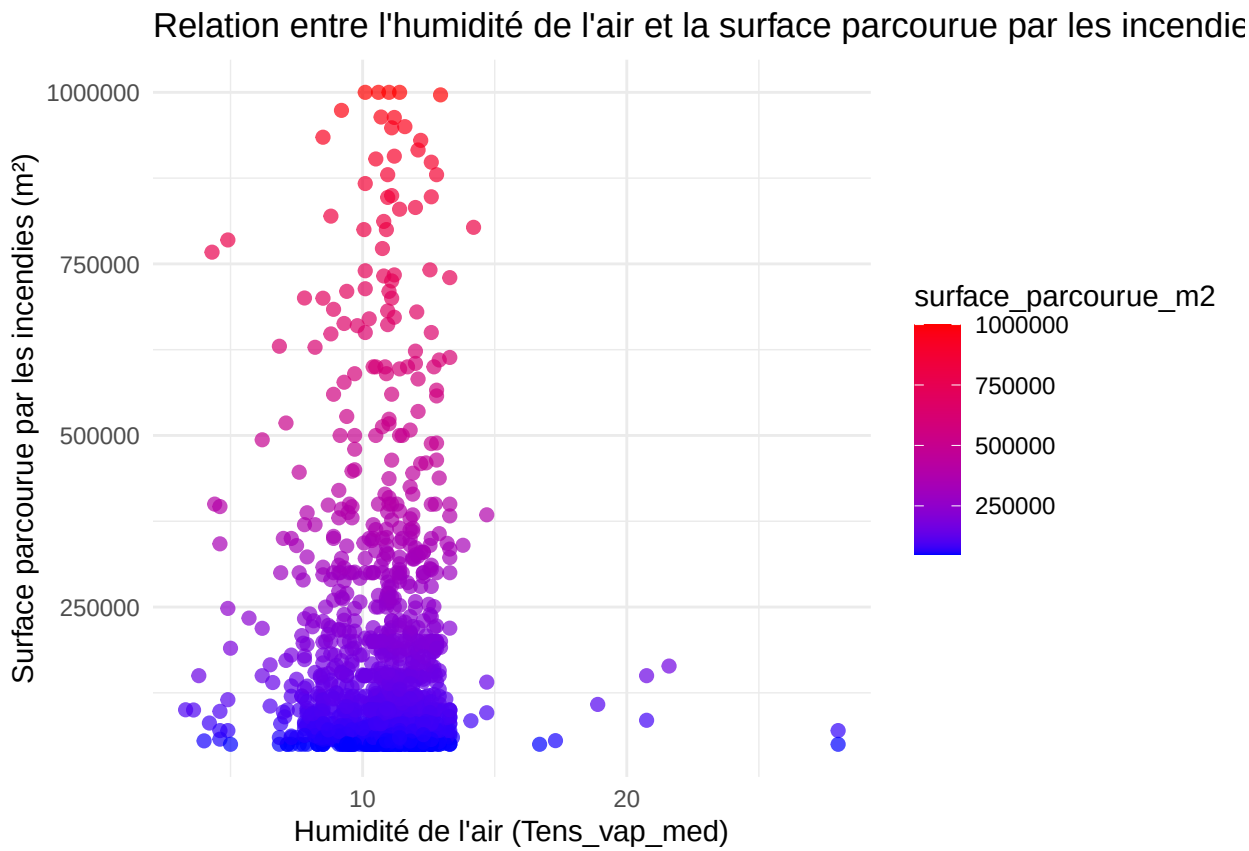
Cet histogramme indiquant une concentration autour de 10-12 hPA suggère un climat tempéré, caractérisé par une humidité généralement modérée la majorité du temps, mais avec des périodes plus humides de manière sporadique.

2. **Diagrame de dispersion**

Ce schéma nous offre la possibilité d'examiner s'il y a une relation entre l'humidité atmosphérique et l'étendue des incendies (autrement dit, la superficie qu'ils couvrent).

L'objectif est d'observer comment l'humidité influe sur la taille des incendies de manière perceptible.

```
library(ggplot2)
data <- read.csv("../Exports/export_Humidites.csv")
ggplot(data, aes(x = Tens_vap_med, y = surface_parcourue_m2)) +
  geom_point(aes(color = surface_parcourue_m2), size = 2, alpha = 0.7) +
  scale_color_gradient(low = "blue", high = "red") +
  labs(title = "Relation entre l'humidité de l'air et la surface parcourue par les incendies",
       x = "Humidité de l'air (Tens_vap_med)",
       y = "Surface parcourue par les incendies (m²)") +
  theme_minimal()
```



Dans ce diagramme, seul l'axe des X représente l'air mesuré par la tension de vapeur moyenne. Les valeurs varient de 0 à 25 hPa. L'axe des Y représente la superficie parcourue par les incendies, exprimée en mètres carrés.

Nous avons choisi d'utiliser un graphique de type nuage de points, également appelé **scatter plot**, où chaque point représente une observation (incendie) associée à deux variables. Premièrement, il s'agit de l'humidité atmosphérique au moment du sinistre, et en second lieu, de la superficie affectée par cet incendie (Y).

La couleur des points varie du bleu au rouge, conformément à l'échelle indiquée sur la droite. Les points bleus représentent des surfaces inférieures à 250 000 m², alors que les points rouges sont associés à des surfaces supérieures.

Débutons par l'étude de la répartition des points, en mettant d'abord l'accent sur leur concentration. La plupart des points se regroupent dans la plage d'humidité de 0 à 15 hPa, avec une densité particulièrement élevée entre 5 et 12 hPa. Cela entraîne des niveaux d'humidité modérés.

On remarque également que le nombre de points au-delà de 20 hPa est très limité, ce qui suggère que les incendies dans des conditions extrêmement humides sont rares.

En ce qui concerne la superficie affectée par les incendies, allant de 0 à 250 000 m², on remarque que la majorité d'entre eux ont une portée assez restreinte. Pour les surfaces allant jusqu'à 1 000 000 m2, le graphique semble indiquer que les incendies de plus grande ampleur sont moins communs.

Concernant le lien entre l'humidité et la surface parcourue, on remarque une concentration de points bleus dans l'intervalle d'humidité de 5 à 15 hPa. Cela indique que les incendies de faible envergure se produisent plus souvent dans des conditions d'humidité modérées. En ce qui concerne les points rouges, ils sont plus éparpillés et se situent dans la même fourchette, bien qu'il existe des points rouges dans des zones où l'humidité est à la fois plus basse et plus élevée.

On ne constate pas de lien clair et direct entre le taux d'humidité de l'air et la superficie touchée par les

feux. Les feux de grande envergure (indiqués par des points rouges) surviennent à divers niveaux d'humidité, toutefois, la plupart d'entre eux (qu'ils soient petits ou grands) se regroupent dans l'intervalle de 5 à 15 hPa.

Cependant, une tendance mineure peut être observée : les incendies de plus grande envergure (près de 1 000 000 m²) ont l'air de survenir un peu plus fréquemment dans des conditions d'humidité plus basse (environ 5 hPa ou moins), où l'air est plus sec, ce qui facilite la diffusion des flammes. Néanmoins, cette tendance n'est pas très prononcée.

3. **Comparaison avec l'histograme precedent:**

L'histogramme précédent nous indiquait que la pression de vapeur moyenne se situait approximativement entre 5 et 15 hPa, avec un sommet autour de 10-12 hPa. Cette distribution est confirmée par ce nuage de points.

Les quelques rares points au-delà de 20 hPa dans l'histogramme témoignent de la confirmation que les incendies sont peu fréquents dans des conditions très humides.

4. **Analyse Statistique**

Après avoir réalisé l'analyse des deux graphiques, nous sommes en mesure d'effectuer une analyse statistique.

Nous allons effectuer un calcul de corrélation.Elle nous offrira la possibilité d'évaluer l'intensité et le sens de la corrélation linéaire entre l'humidité atmosphérique et les dimensions des feux.

Avant tout, définissons les choses. Il s'agit d'une mesure statistique qui illustre la force et la direction d'un lien entre deux variables. Elle nous aide, de manière simple, à saisir comment deux variables se déplacent l'une par rapport à l'autre. On utilise le coefficient de corrélation de Pearson, qui se situe entre -1 et 1, pour quantifier la corrélation. Avec 1 représentant une corrélation parfaitement positive, -1 une corrélation parfaitement négative et 0 signifiant aucune corrélation.

```
correlation <- cor(data$Tens_vap_med, data$surface_parcourue_m2, use = "complete.obs")
print(paste("Corrélation entre Tens_vap_med et surface_parcourue_m2: ", correlation))
```

```
## [1] "Corrélation entre Tens_vap_med et surface_parcourue_m2:  -0.0211442157372533"
```

Dans ce code, nous avons fait appel à la fonction « cor() » afin de déterminer la corrélation de Pearson entre Tens_vap_med et surface_parcourue_m2, en omettant les variables manquantes. La méthode cor() nous donnera un coefficient de corrélation, comme expliqué précédemment.

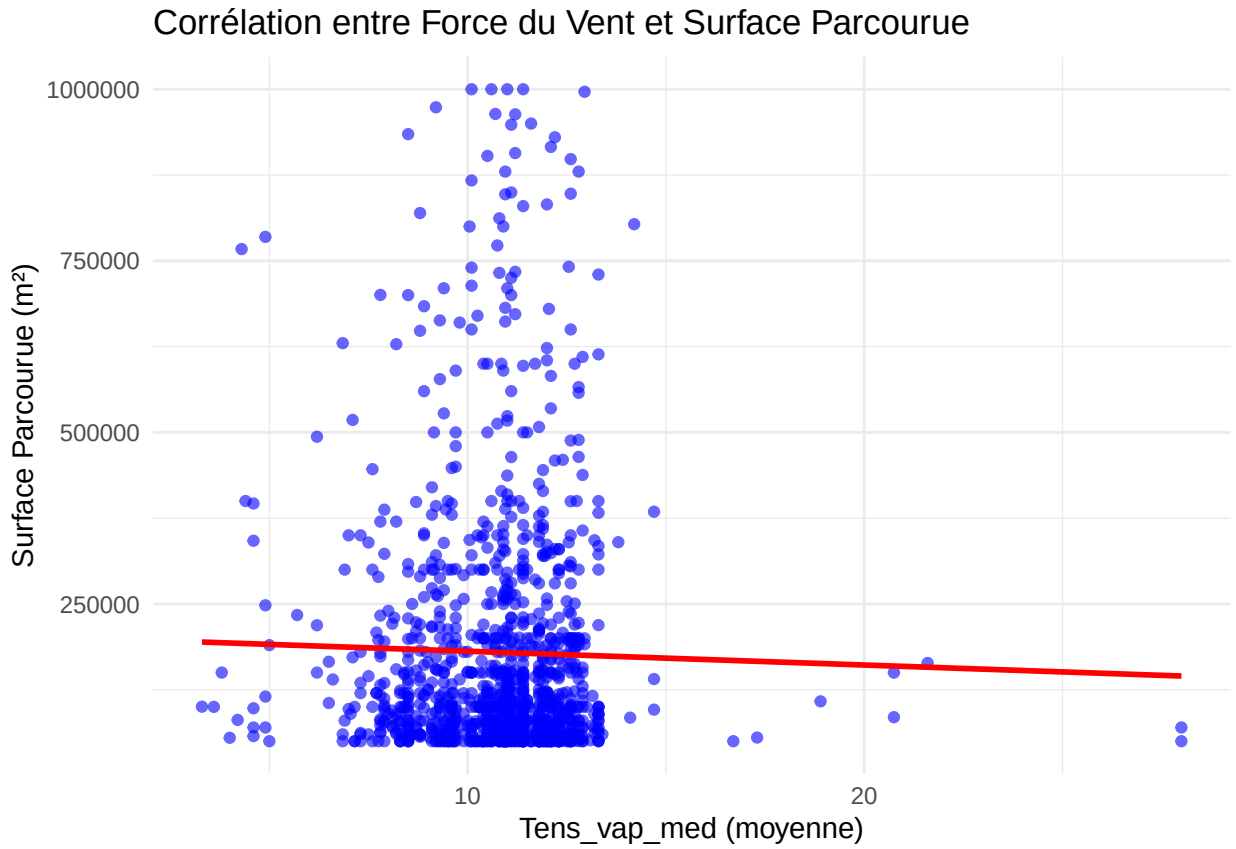
1. Si la corrélation est haute, proche de 1, cela nous permet d'affirmer qu'une hausse de l'humidité est liée à une extension des incendies.
2. Si la corrélation est négative (proche de -1), cela implique qu'une hausse de l'humidité est liée à une réduction de la grandeur des feux.
3. Une corrélation proche de 0 suggère l'absence d'une relation linéaire manifeste.

L'analyse de corrélation révèle qu'il n'existe pas de lien linéaire prononcé entre le taux d'humidité et l'ampleur des feux dans vos informations. Selon cette étude, l'humidité semble avoir une influence marginale sur l'ampleur des incendies.

Apres avoir calcule la correlation on va y tracer le graphe de cette correlation

```r
library(ggplot2)
data <- read.csv("../Exports/export_Humidites.csv")
ggplot(data, aes(x = Tens_vap_med, y = surface_parcourue_m2)) +
  geom_point(color = "blue", alpha = 0.6) +  # Ajoute les points
  geom_smooth(method = "lm", color = "red", se = FALSE) +  # Ajoute la droite de régression
  labs(title = "Corrélation entre Force du Vent et Surface Parcourue",
       x = "Tens_vap_med (moyenne)",
       y = "Surface Parcourue (m²)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



5.3.2.3 Relation entre le vent et la propagation des incendies Pour examiner la question relative à la corrélation entre le vent et la diffusion des incendies sur le sol français, nous adopterons une approche distincte en segmentant ce sujet en quatres sous-questions.

C'est pourquoi, pour aborder ce problème, on devrait considérer ces cinq enjeux :

1. Force du vent
2. Surface parcourue par le feu en fonction de la Force du vent
3. Force du vent par zone géographique
4. Surface parcourue par le feu par zone géographique

Avant d'examiner les quatre sous-problèmes, nous allons inspecter la corrélation.Nous pourrions vérifier la corrélation entre la puissance du vent et la surface parcourue en mètres carrés. En procédant ainsi, nous pourrions déterminer s'il existe une relation linéaire entre ces deux variables.

On remarque donc que si le coefficient de corrélation se rapproche de -1 ou 1, cela signifie qu'il existe une relation linéaire significative entre ces deux variables.

```
data <- read.csv("../Exports/export_vents.csv")
cor(data$Force_vent_med, data$surface_parcourue_m2, use = "complete.obs")
```

```
## [1] 0.03137438
```

Puisque le coefficient de corrélation de Pearson est proche de 0, cela indique qu'il n'existe pas de relation linéaire significative entre les deux variables.

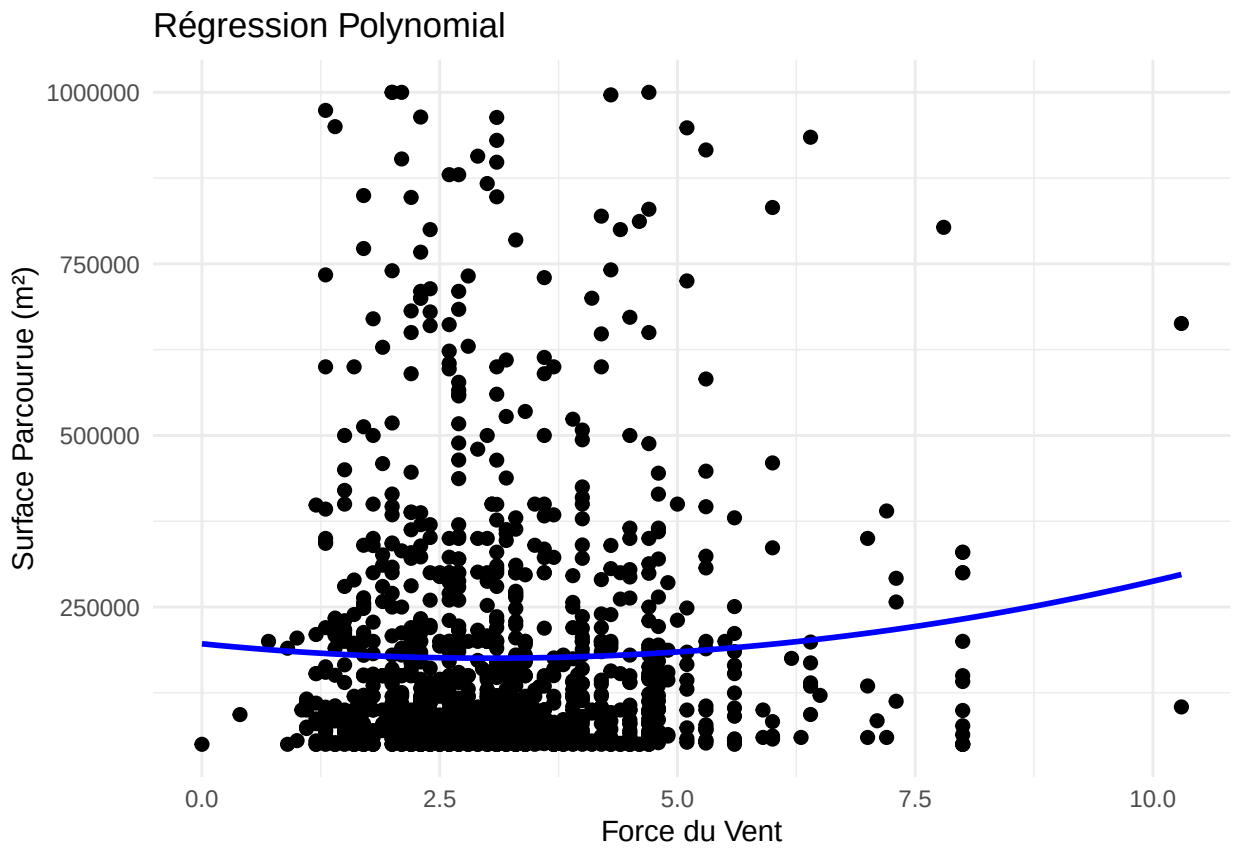
Pour une perspective plus statistique, nous allons essayer la régression polynomiale.C'est une technique employée pour modéliser le lien entre une variable indépendante, comme dans notre situation la puissance du véhicule, et une variable dépendante, ici la distance parcourue par le feu, à condition que ce lien ne soit pas linéaire. Cela s'applique à notre situation.

Cela nous offre la possibilité de mieux saisir les courbures ou les tendances complexes des données, contrairement à une régression linéaire simple qui postule une relation proportionnelle constante entre les deux variables.

Dans ce contexte, nous employons une régression polynomiale de degré 2, que l'on peut également qualifier de régression quadratique.Cette dernière intègre finalement le terme linéaire (force du vent) ainsi que le terme quadratique (force du vent au carré).

On va détailler davantage pourquoi nous avons opté pour une régression polynomiale, étant donné qu'elle nous offre la possibilité de saisir des relations non linéaires entre les variables.Une condition est établie si le lien entre la puissance du vent et la superficie touchée par le feu suit une courbe, comme c'est le cas dans notre scénario, avec une accélération rapide initiale qui se ralentit à mesure que la force du vent s'intensifie.Dans ce contexte, une régression polynomiale serait plus appropriée qu'une régression linéaire simple.

```
library(ggplot2)

# Charger les données
data <- read.csv("../Exports/export_vents.csv")

# Créer un graphique avec régression polynomial
ggplot(data, aes(x = Force_vent_med, y = surface_parcourue_m2)) +
  geom_point(color = "black", size = 2) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), color = "blue", se = FALSE) +
  labs(title = "Régression Polynomial",
       x = "Force du Vent",
       y = "Surface Parcourue (m²)") +
  theme_minimal()
```



Nous allons détailler, à travers une analyse informatique, la manière dont nous avons réalisé la régression polynomiale.Pour cela, nous avons utilisé la méthode **geom_smooth()**, accompagnée de l'argument **method = "lm"** qui indique que nous sommes en présence d'un modèle de régression.La formule = y ~ poly(x, 2) précise qu'il s'agit d'une régression polynomiale de degré 2 (quadratique).

Nous allons pouvoir intégrer une courbe de régression linéaire grâce à la méthode **geom_smooth(method = "lm")**. L'indication précise que le modèle repose sur une analyse de régression linéaire.

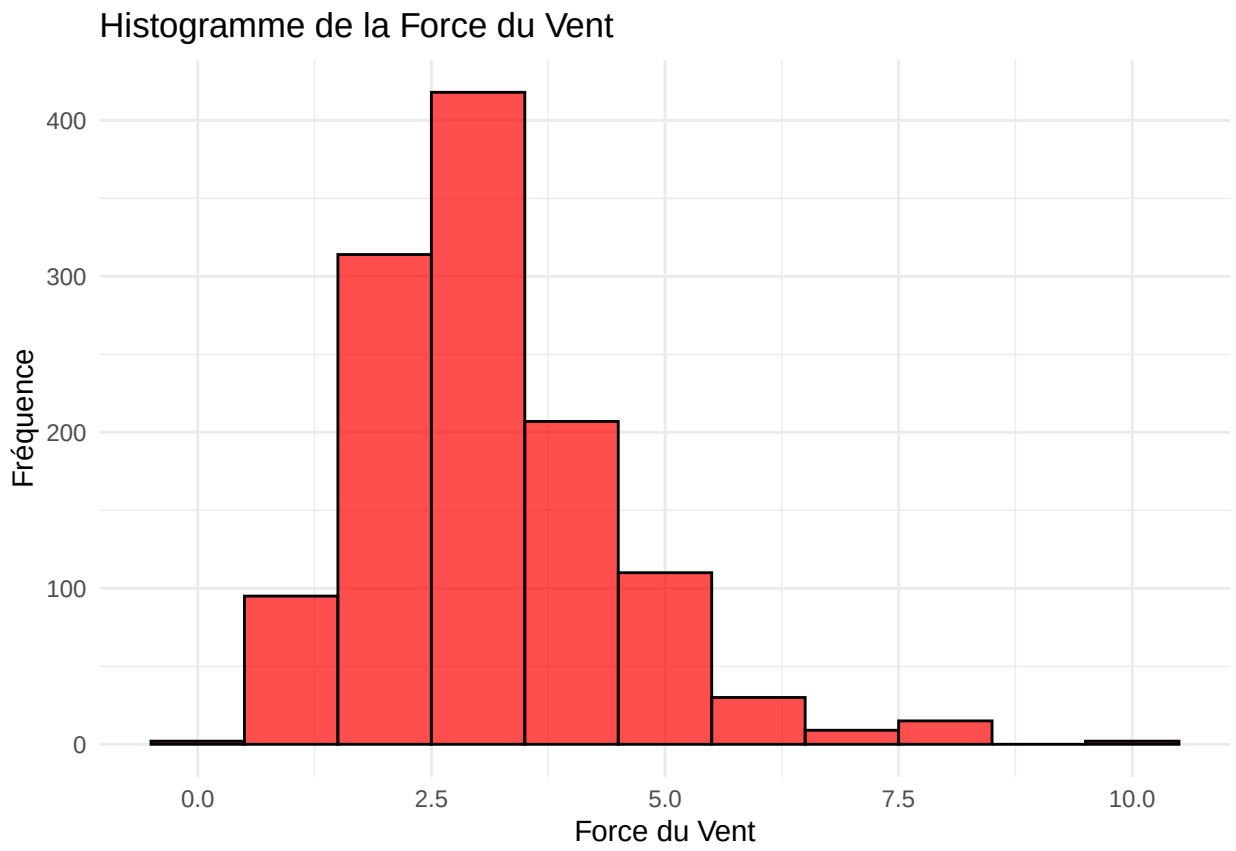
Cette spécification, **formula = y ~ poly(x, 2)**, indique au langage R que nous souhaitons un modèle polynomial du second degré. **poly(x, 2)** est une fonction qui produit les termes x et x^2, représentant respectivement la vitesse du vent et son carré.Cela permettra à la régression de s'ajuster à la fois à une pente linéaire et à une courbure quadratique des données.

L'attribut **color="blue"** nous offre la possibilité de peindre la courbe de régression en bleu, ce qui permet une distinction claire avec les autres éléments du graphe. Et en outre, **se=FALSE** nous permettra de désactiver l'affichage de la marge d'erreur autour de la courbe de régression.

1. **Force du vent**

Dans cette sous-problematique on va analyser la force du vent:

```
library(ggplot2)
data <- read.csv("../Exports/export_vents.csv")
ggplot(data, aes(x = Force_vent_med)) +
  geom_histogram(binwidth = 1, fill = "red", color = "black", alpha = 0.7) +
  labs(title = "Histogramme de la Force du Vent", x = "Force du Vent", y = "Fréquence") +
  theme_minimal()
```



1. **Analyse Informatique:**

Tout d'abord, nous allons détailler la méthode de construction de notre graphique. Nous avons utilisé la bibliothèque ggplot du langage R pour élaborer cet histogramme.Ensuite, nous avons importé nos données en spécifiant le chemin relatif du fichier CSV que nous avons conçu et développé à l'aide du langage Python et de la gestion des fichiers.

Dans le processus de chargement du fichier, nous avons employé la méthode read.csv pour interpréter le fichier CSV, et nous avons enregistré ces informations dans la variable data.

Par la suite, nous avons fait appel à la méthode, définie dans la bibliothèque ggplot2, soit la méthode ggplot. Nous lui avons précisé l'emplacement des données stockées et mis en place une autre méthode aes, signifiant **aesthetics**, pour indiquer quelles colonnes de données devraient être visualisées sur les axes. Dans notre situation, nous attribuons l'axe des x à la variable Force_vent_med qui va symboliser l'intensité du vent mesurée.

Par la suite, nous employons la méthode **geom_histogram()** pour intégrer l'histogramme à la représentation graphique. Nous spécifions la largeur des barres, la teinte que nous voulons utiliser pour leur remplissage, la nuance du contour et également le degré de transparence des barres.

On termine par la définition des étiquettes et des titres de notre histogramme grâce à la méthode **labs()**. On y inclut le titre du graphique ainsi que les libellés pour les deux axes, x et y. De plus, comme nous avons opté pour un style minimaliste, nous avons utilisé la méthode **theme_minimal()** afin de conférer un aspect plus contemporain au graphique.

2. **Analyse Statistique:**

On présume que la force du vent est évaluée en kilomètres par heure.

Cette histograme represente la distribution de la force du vent mesure en Km/h. L'échelle horizontale montre la puissance du vent, avec des valeurs variant de 0 à 10,5.L'axe des y symbolise la fréquence, soit le nombre d'apparitions de chaque plage de force du vent, avec des valeurs se situant approximativement entre 0 et 400.

L'histogramme indique une distribution asymétrique vers la droite, ce qui implique une concentration accrue de données pour des valeurs faibles de la force du vent, avec une longue traîne vers les valeurs plus élevées.

Le sommet de l'histogramme se trouve approximativement dans la gamme de 3.5 à 4.0 pour la force du vent, avec une fréquence avoisinant les 400. Cela nous indique que dans cet échantillon, la force du vent la plus courante se situe dans cette plage.

Concernant la portée des valeurs, la force du vent varie approximativement de 0 à 10,5. Toutefois, les occurrences de valeurs dépassant 8 sont extrêmement rares, ce qui indique que de très forts vents sont peu fréquents dans cet ensemble de données.

Presque 80% des données se regroupent entre 0 et 5.5 environ, ce qui indique que cet échantillon est principalement dominé par les vents légers à modérés.

La moyenne de cet histogramme se situe légèrement au-dessus du mode, estimée approximativement entre 4.0 et 4.5.Concernant la médiane, elle se positionne probablement autour de 3.5, car la distribution penche vers la droite, divisant l'échantillon en deux segments égaux.

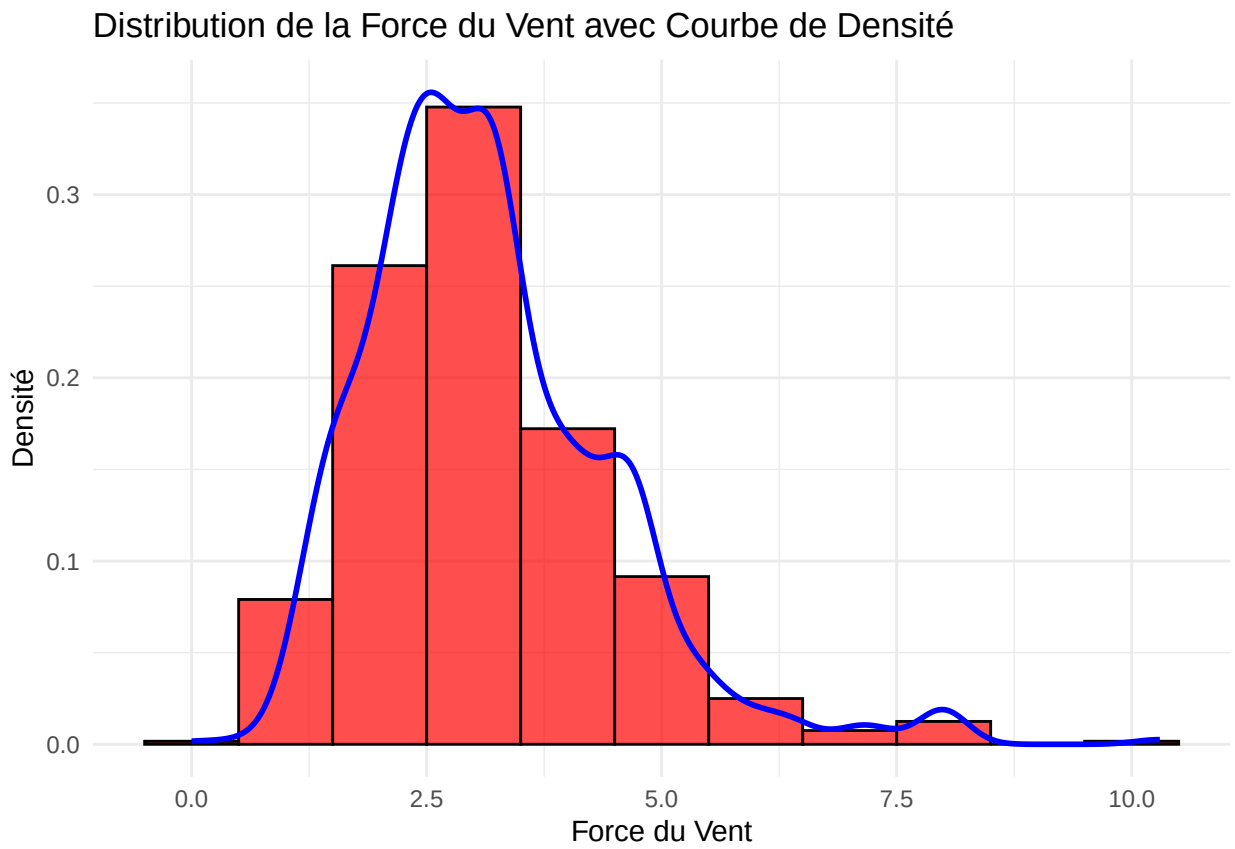
Cet histogramme pourrait illustrer des relevés de la puissance du vent sur une période déterminée.La prévalence de brises légères à modérées (0.0 à 5.5) indique un climat plutôt paisible, avec la présence exceptionnelle de vents puissants au-delà de 8, qui pourraient survenir lors d'événements rares associés à des phénomènes météorologiques tels que des tempêtes.

Maintenant on va tracer une courbe de densite:

```
# Histogramme avec courbe de densité pour la Force du Vent
data <- read.csv("../Exports/export_vents.csv")
ggplot(data, aes(x = Force_vent_med)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill = "red", color = "black", alpha = 0.7) +
  geom_density(color = "blue", size = 1) +
  labs(title = "Distribution de la Force du Vent avec Courbe de Densité", x = "Force du Vent", y = "Dens
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Dans ce graphique, nous avons tracé une courbe de densité pour illustrer la répartition de la puissance du vent dans nos données.Ainsi, l'axe des X reflète la force du vent tandis que l'axe des Y indique la densité, c'est-à-dire en d'autres termes,la fréquence relative plutôt que le nombre absolu.

La densité sert à normaliser l'histogramme afin qu'il présente une surface totale de 1, ce qui simplifie la comparaison entre différentes distributions ayant des échantillons variés.

La courbe de densité représente une estimation non paramétrique de la fonction de densité des probabilités. Elle est fluide et représente la probabilité que les observations prennent une valeur dans une certaine plage de force du vent.

La courbe de densité fournit une indication sur la structure globale de la distribution des données.

Si la courbe est centrée sur une certaine valeur, cela signifie que la plupart des données sont groupées autour de cette valeur.

Si la courbe est plate ou étendue, cela suggère une plus grande variabilité des données.

Nous allons également décrire la manière dont nous avons réalisé la partie informatique en utilisant la fonction **geom_histogram(aes(y = ..density..))**, ce qui nous a permis de construire l'histogramme. L'argument **aes(y = ..density..)** spécifie que l'axe des ordonnées doit refléter la densité plutôt que la fréquence brute. Cela nous permettra de standardiser les données afin que l'aire totale sous l'histogramme soit de 1.

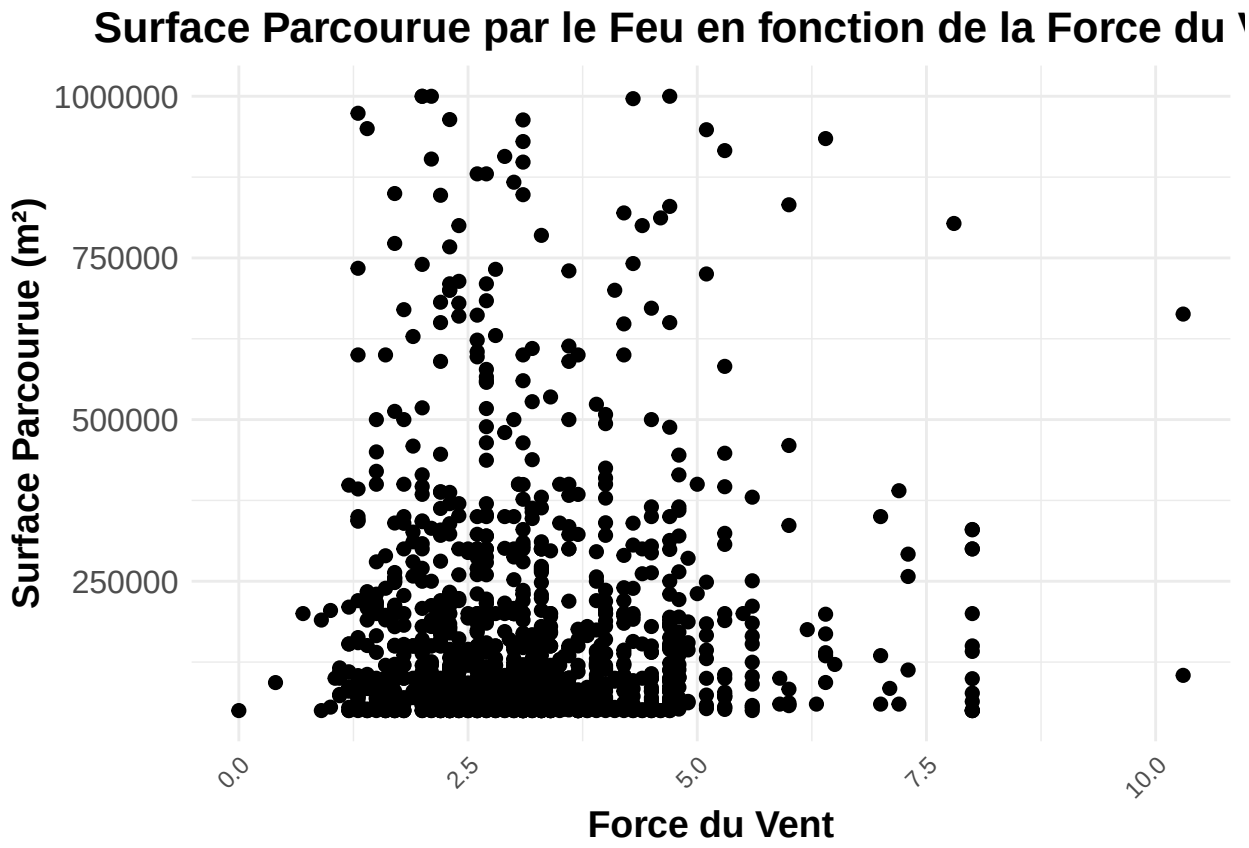
La commande **geom_density(color = "blue", size = 1)** nous offre la possibilité d'intégrer la courbe de densité.

La fonction **labs(title = "Distribution de la Force du Vent avec Courbe de Densité", x = "Force du Vent", y = "Densité")** nous donne la possibilité de spécifier le titre du diagramme ainsi que les étiquettes des axes X et Y.

La fonction **theme_minimal()** applique un style épuré au graphique, éliminant les éléments visuels superflus.

2. **Surface parcourue par le feu en fonction de la force du vent**

```r
library(ggplot2)

data <- read.csv("../Exports/export_vents.csv")

ggplot(data, aes(x = Force_vent_med, y = surface_parcourue_m2)) +
  geom_point(color = "black", size = 2) +
  labs(title = "Surface Parcourue par le Feu en fonction de la Force du Vent",
       x = "Force du Vent",
       y = "Surface Parcourue (m²)") +
  theme_minimal() +  # Thème minimal pour une présentation épurée
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.text.y = element_text(size = 12),
    axis.title = element_text(size = 14, face = "bold"),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5)
  )
```



1. **Analyse Informatique:**

Tout d'abord, nous allons décrire comment nous avons développé notre code. Pour commencer, nous avons importé notre bibliothèque ggplot2 afin de créer des visualisations complexes en langage R. Par la suite, nous

avons chargé notre fichier contenant les données grâce à la méthode **read.csv()** et avons assigné le résultat à une variable nommée data. Cette variable data représente l'objet où nous conservons ces informations.

Ensuite, nous élaborons notre diagramme de dispersion en utilisant la méthode **ggplot()**, qui intègre les données. Par ailleurs, l'axe des x, déterminé par l'aes, représentera la force du vent alors que l'axe des y illustrera la surface parcourue en m^2.

Actuellement, nous sommes dans la phase où nous ajoutons des points pour pouvoir représenter les données. Nous utilisons la méthode **geom_point()** pour ce faire. Nous avons spécifié la couleur des points en noir et aussi déterminé la taille des points pour une visibilité optimale en utilisant l'option size = 2.

Pour une lecture optimale du graphique, nous avons utilisé la fonction **labs()**. Nous avons ajouté le titre au graphique et précisé les noms des deux axes en recourant à x et y.

Comme le graphe precedent on a utilise la methode du **theme_minimal()** pour le mettre dans un desin plus minimaliste

À l'instar du graphique précédent, nous avons employé la technique du **theme_minimal()** afin de le présenter dans une design plus épuré.

Pour accroître la clarté des axes et du titre, nous faisons appel à la fonction **theme()**. En manipulant **axis.text.x** et **axis.text.y**, nous avons employé l'option **element_text** pour le rendre incliné, modifier sa taille, etc. Quant aux attributs des deux axes, nous avons utilisé axis.title pour mettre le texte en gras grâce à l'attribut **face**.Pour le titre principal du graphique, on utilise **plot.title** pour ajuster ses caractéristiques à l'aide de **size**, **face** et **hjust**.L'utilisation de **hjust** permet de positionner le titre au centre.

2. **Analyse Statistique:**

Commençons par examiner ce sous-problème qui se focalise sur la superficie que couvre le feu en relation avec l'intensité du vent. Tout d'abord, on remarque que la force du vent varie entre 0 et 10 unités. De plus, la superficie affectée par le feu ou les incendies est presque de 0 à 1 000 000.

On remarque une concentration significative de points entre 0 et 5 unités de force du vent, avec une superficie allant de 0 à environ 750 000 m^2.

Il est également possible de conclure qu'au-delà de cinq unités, les points commencent à devenir rares et la surface parcourue tend à se stabiliser tout en réduisant légèrement.

L'accumulation de points indique qu'une intensification du vent a tendance à élargir la portée du feu, particulièrement dans le cas de vents faibles à modérés.Toutefois, ce lien n'est pas rigoureusement linéaire, étant donné la large répartition des points.

Au-delà d'une certaine intensité de vent (approximativement 7-8 unités), la surface couverte ne paraît pas s'accroître de manière proportionnelle, ce qui pourrait suggérer un phénomène de saturation ou des éléments restrictifs (tels que la disponibilité du combustible, l'humidité ou la topographie).

Il est également important de souligner que la puissance du vent joue un rôle crucial dans la diffusion des incendies, puisqu'elle transporte de l'oxygène et les particules embrasées amplifient ainsi la vitesse de propagation et l'étendue touchée. Ceci justifie l'orientation initiale à la hausse.

Synthese de la Problematique

Ainsi, pour résumer cette question, nous avons d'abord établi une corrélation entre la force du vent et la superficie touchée par le feu.Nous avons constaté qu'il n'y a pas de relation linéaire entre ces deux facteurs. Pour saisir plus fidèlement la complexité de la relation, nous avons réalisé une régression polynomiale de degré 2. Cette méthode nous a autorisé à noter que la surface affectée par l'incendie a tendance à croître plus significativement à des niveaux faibles de la force du vent avant de diminuer à mesure que cette dernière devient plus intense.Cela indique qu'il y a une dynamique non linéaire entre ces deux éléments,avec un effet d'accélération initial suivi d'un processus de plafonnement.

L'analyse effectuée sur l'histogramme et la courbe de densité a montré que la majorité des forces du vent se concentre autour de valeurs basses à modérées, avec une queue notable vers les valeurs plus hautes.Cela indique que, dans notre échantillon,les vents de faible à moyenne intensité sont prédominants, tandis que les vents de très haute intensité sont plutôt rares.

Le diagramme de dispersion a révélé une tendance à l'augmentation de la superficie brûlée avec la puissance du vent, bien que cette corrélation ne soit pas entièrement linéaire. Les informations suggèrent que, face à des vents de faible à modéré (jusqu'à environ 5 unités de force),la superficie touchée par le feu s'accroît rapidement.Toutefois, au-delà d'un certain seuil de force du vent, l'expansion de la zone parcourue semble se stabiliser, indiquant d'autres facteurs limitants que'il ne s'agit pas seulement du vent, comme la disponibilité du carburant ou la configuration du terrain.

Les données recueillies indiquent que même si la puissance du vent a un impact significatif sur l'expansion des incendies, elle ne constitue pas le seul élément décisif.D'autres facteurs tels que la disponibilité des carburants, l'humidité du sol et la topographie locale ont aussi un impact sur la diffusion du feu.Ainsi, même si une augmentation de la vitesse du vent peut effectivement favoriser la diffusion des incendies, cette corrélation devient moins directe et plus complexe lorsque le vent atteint des intensités plus extrêmes.

5.3.2.4 Impact des conditions météorologiques extrêmes sur les incendies

5.3.2.5 Effet des radiations solaires sur les incendies

5.3.2.6 Température et nombre d'incendies par période de la journée Pour pouvoir analyser notre problemarique qui se concerne sur la temperature et le nombre des incendies par periode journee on va diviser notre problematique en des sous problematiques:

- Relation entre la température maximale quotidienne et le nombre d'incendies
- Impact combiné de la température et de l'heure de la journée sur les incendies
- Seuils de température critique pour l'émergence des incendies

1- **Relation entre la température maximale quotidienne et le nombre d'incendies:**

```
library(ggplot2)
library(dplyr)
agg_data <- read.csv("../Exports/export_incendiestempheure.csv")

# Calculer le nombre d'incendies par jour (regroupement par an, mois, jour)
agg_data_daily <- agg_data %>%
  group_by(annee, mois, jour) %>%
  summarise(nb_incendies = n(),   # Nombre d'incendies par jour
            tmax_med = mean(tmax_med, na.rm = TRUE))   # Température maximale moyenne par jour
```

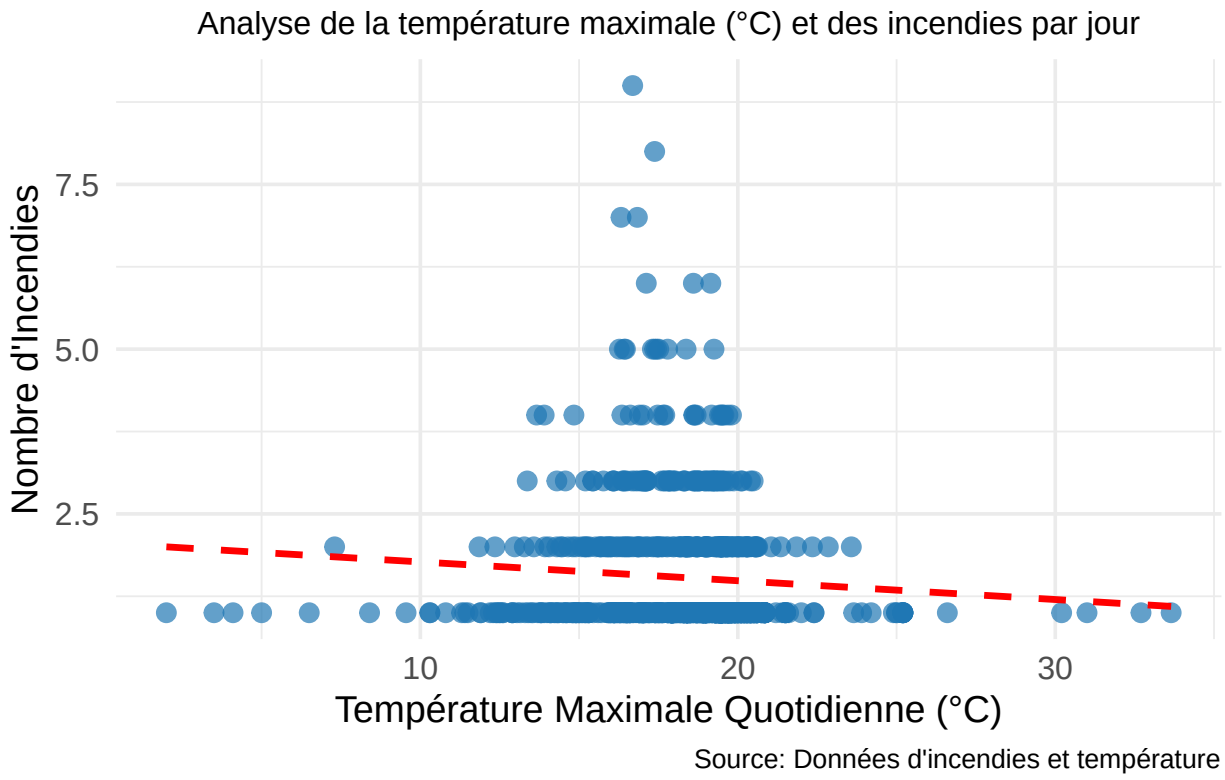
```
## `summarise()` has grouped output by 'annee', 'mois'. You can override using the
## `.groups` argument.
```

```
# Créer un graphique en nuage de points avec ggplot2
ggplot(agg_data_daily, aes(x = tmax_med, y = nb_incendies)) +
  geom_point(color = "#1f77b4", size = 3, alpha = 0.7) +  # Points de données (couleur et opacité ajust
  geom_smooth(method = "lm", color = "red", se = FALSE, linetype = "dashed", size = 1.2) +  # Ligne de
  labs(
    title = "Relation entre la température maximale quotidienne et le nombre d'incendies",
```

```r
    subtitle = "Analyse de la température maximale (°C) et des incendies par jour",
    x = "Température Maximale Quotidienne (°C)",
    y = "Nombre d'Incendies",
    caption = "Source: Données d'incendies et température"
  ) +
  theme_minimal(base_size = 14) +  # Thème minimal avec taille de base augmentée pour meilleure lisibil
  theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),  # Centrer et formater le titre
    plot.subtitle = element_text(hjust = 0.5, size = 12),               # Centrer et formater le sous-t
    axis.title = element_text(size = 14),                               # Taille des titres des axes
    axis.text = element_text(size = 12),                                # Taille des labels des axes
    plot.caption = element_text(size = 10, hjust = 1)                   # Taille et alignement de la l
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

entre la température maximale quotidienne et le nombr



Analyse Informatique

Pour réaliser notre graphique, nous avons eu recours à deux bibliothèques : la bibliothèque **ggplot2**, qui nous offre la possibilité de créer des représentations graphiques en utilisant le langage R, et **dplyr**.Cette bibliothèque est intégrée au **tidyverse**, facilitant une manipulation efficace des données.Elle offre la possibilité de sélectionner, filtrer, grouper et résumer les données de manière performante.

Par la suite, nous utilisons la méthode **read.csv** pour charger les données contenues dans le fichier CSV dans une variable nommée **agg_data**.

Par la suite, on effectue les statistiques journalières %>% en utilisant l'opérateur pipe de la bibliothèque **dplyr**, ce qui nous permet d'enchaîner plusieurs opérations. Ensuite, nous utilisons la fonction **groupby()** qui nous autorise à regrouper les données par année, mois et jour.

Ceci nous autorisera à effectuer des calculs statistiques sur une base quotidienne. En outre, la technique **summarise** nous offre la possibilité d'effectuer des résumés statistiques grâce à l'utilisation de la méthode **n()**, qui permet de déterminer le nombre d'incendies par jour en comptant le nombre d'observations pour chaque groupe, soit ici, chaque jour.

En outre, on fait appel à la technique **mean** pour déterminer la température maximale moyenne quotidienne en omettant les valeurs absentes grâce à **na.rm = True**.

Après avoir effectué le calcul des statistiques journalières, nous créons le graphique en employant la fonction **ggplot()** de la librairie **ggplot2**.

Nous l'élaborons en utilisant les données regroupées de la variable contenant les informations. Grâce à la méthode **aes()**, nous définissons les axes.

Par la suite, nous matérialisons les points de données sous forme de nuages de points grâce à l'usage du procédé **geom_point()**.

Après l'utilisation de **geom_smooth()**, on superpose à ce graphique une ligne de régression linéaire pour mettre en évidence le lien entre la température et le nombre d'incendies, en précisant l'option method="lm" pour indiquer la méthode utilisée afin d'ajouter une régression linéaire.

La fonction labs() est employée pour définir les titres des axes, tandis que **theme_minimal()** est utilisée pour appliquer un style minimaliste.

Analyse Statistique:

On précise que l'axe horizontal indique la température maximale variant de 0 à 35 degrés et l'axe vertical dénote le nombre d'incendies oscillant entre 0 et 7,5.

Par ailleurs, on note que les points bleus symbolisent les observations.

En d'autres termes, chaque point représente un jour avec une température maximale spécifique et un nombre précis d'incendies.

Dans une échelle de 0 à 7,5, on détermine que les points bleus illustrent les observations. En d'autres termes, chaque point correspond à un jour donné avec une température maximale spécifique et un nombre déterminé d'incendies.

Il s'agit probablement d'une régression qui trace la relation moyenne entre la température et le nombre d'incendies.

Bien que les points soient éparpillés, on remarque une concentration plus importante de données entre 15 et 25 degrés, avec une prédominance d'incendies se situant généralement entre 0 et 3

On note des températures plus extrêmes, mais les données sont moins abondantes.

Dans notre graphique, la courbe rouge indique une corrélation non linéaire entre la température maximale et le nombre d'incendies.

De zéro à environ quinze, le nombre d'incendies paraît légèrement en hausse, mais il reste faible, oscillant entre 1 et 2 par jour en moyenne.

Entre 15 et 25, la courbe connaît un sommet avec une moyenne de 2 à 3 incendies.

C'est dans cette fourchette de températures que l'on observe le plus grand nombre d'incendies.

Au-delà de 25 degrés, la courbe diminue progressivement, signalant une réduction du nombre moyen d'incendies lorsque les températures excèdent 25 degrés.

La dispersion des points autour de la courbe de régression indique une variabilité importante, comme on peut le voir dans notre graphique où à 20 degrés le nombre d'incendies fluctue entre 0 et 5.

Cette fluctuation indique que la température seule ne peut rendre compte du nombre d'incendies.

Le lien constaté est en accord avec nos connaissances sur les feux de forêt : des températures plutôt élevées (15-25 °C) peuvent contribuer à l'éclosion de ces derniers en déshydratant la végétation, particulièrement si elles sont associées à une faible humidité ou à des vents.

Toutefois, dans des conditions de chaleur extrême (supérieure à 25 °C), d'autres paramètres tels qu'une diminution de l'humidité ou des actions préventives (interdiction de feux, vigilance renforcée) pourraient faire baisser le nombre d'incendies.

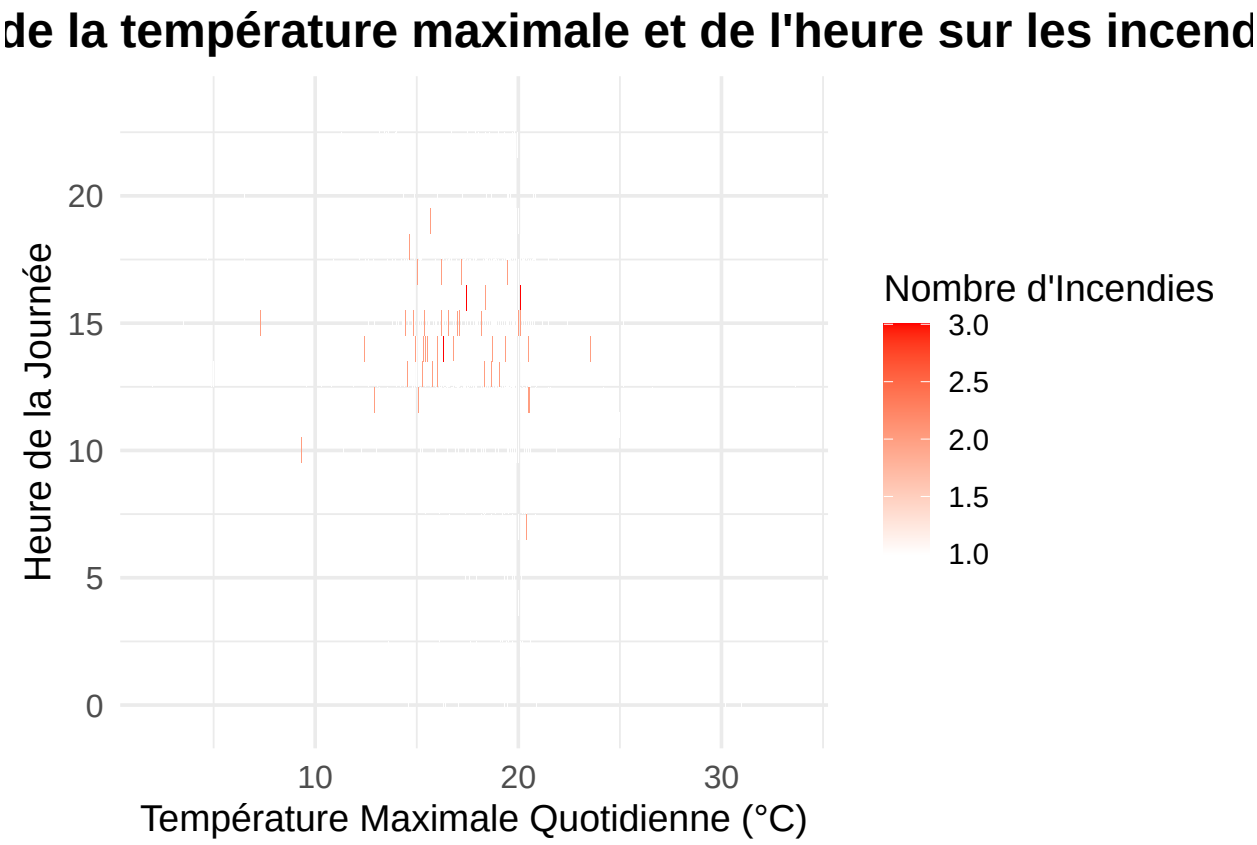
Il est compréhensible que le nombre d'incendies à des températures extrêmement basses (proches de 0 °C) soit réduit, étant donné que les conditions ne favorisent pas la dissémination des flammes (végétation humide, gel, etc.).

2- **Impact combiné de la température et de l'heure de la journée sur les incendies**

```r
library(ggplot2)
library(dplyr)
agg_data <- read.csv("../Exports/export_incendiestempheure.csv")

# Aggréger les données pour obtenir le nombre d'incendies, la température et l'heure
agg_data_combined <- agg_data %>%
  group_by(annee, mois, jour, heure) %>%
  summarise(
    nb_incendies = n(),
    tmax_med = mean(tmax_med, na.rm = TRUE)
  )
```

```
## `summarise()` has grouped output by 'annee', 'mois', 'jour'. You can override
## using the `.groups` argument.
```

```r
# Créer un graphique de chaleur (heatmap) pour l'impact combiné de la température et de l'heure
ggplot(agg_data_combined, aes(x = tmax_med, y = heure, fill = nb_incendies)) +
  geom_tile() +  # Créer une carte de chaleur (tiles)
  scale_fill_gradient(low = "white", high = "red") +  # Définir les couleurs de la carte de chaleur
  labs(
    title = "Impact combiné de la température maximale et de l'heure sur les incendies",
    x = "Température Maximale Quotidienne (°C)",
    y = "Heure de la Journée",
    fill = "Nombre d'Incendies"
  ) +
  theme_minimal(base_size = 14) +  # Appliquer un thème minimal
  theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),  # Centrer et formater le titre
    axis.title = element_text(size = 14),                               # Taille des titres des axes
    axis.text = element_text(size = 12),                                # Taille des labels des axes
    plot.caption = element_text(size = 10, hjust = 1)                   # Taille et alignement de la lé
  )
```



Analyse Informatique:

Pour réaliser ce diagramme, deux bibliothèques majeures ont été employées, à savoir **ggplot2** et **dplyr**. Les données ont été importées à l'aide d'un DataFrame, puis nous avons consolidé ces données en utilisant l'opérateur **pipe**, qui nous permet d'enchaîner les opérations de manière claire et cohérente.

Par la suite, nous les organisons en utilisant la méthode **group_by()**, ce qui nous permettra de les réagencer par année, mois, jour et heure.

Cela nous donnera la possibilité de calculer des statistiques spécifiques pour chaque combinaison de ces périodes.

On a également employé la méthode **summarise** et la fonction **n()** pour déterminer le nombre d'incendies pour chaque combinaison d'année, mois, jour et heure.

Cette technique va dénombrer les éléments présents dans chaque groupe. Par la suite, on détermine la température moyenne maximale pour chaque groupe sur une base quotidienne et horaire.

L'argument **na.Rm = True** nous autorise à faire abstraction des valeurs absentes lors de la réalisation du calcul.

Par la suite, on procède à l'élaboration du graphique en employant la méthode **ggplot()** qui permettra de le réaliser en se basant sur les données agrégées. On spécifie les variables à représenter sur les deux axes grâce à la méthode **aes()**.

Suite à cela, **geom_title()** nous offre la possibilité de construire la carte où chaque cellule représente une combinaison d'heure et de température.

Chaque case comportant un grand nombre d'incendies sera teintée en rouge.

Suite à cela, la méthode **scale_fill_gradient** nous autorisera à établir une palette de couleurs employée dans le diagramme de chaleur.

Les tuiles avec un faible nombre d'incendies seront en couleur blanche pour indiquer une faible occurrence d'incendies, tandis que celles avec un nombre élevé d'incendies seront en rouge.

La fonction labs() est employée pour définir les titres des axes, tandis que **theme_minimal()** est utilisée pour appliquer un style minimaliste.

Analyse Statistique:

En observant ce graphique, on détermine la température maximale de chaque jour à l'aide de l'axe horizontal, tandis que l'axe vertical indique l'heure de la journée.

Les points dépeignent les incendies avec une couleur qui indique le volume des incendies en fonction de l'échelle chromatique.

Notre gamme de couleurs va du rose clair 1.0 au rouge sombre 3.0.

Ainsi, chaque point représente un incendie qui s'est produit à une température maximale et à un moment précis, avec une intensité (indiquée par la couleur) correspondant au nombre d'incendies.

Bien que les points sur le graphique soient éparpillés, on remarque une densité plus importante dans certaines zones. La plupart des incendies se regroupent entre 15 et 25, ce qui a du sens. On peut également noter que les incendidés ont tendance à se déclencher surtout entre 10 et 18 heures, avec un pic visible autour de midi à seize heures.

On constate très peu de points hors de ces plages, par exemple, il y a rarement des incendies avant 5h ou après 18h, et très peu à des températures inférieures à 10 ou supérieures à 30 degrés.

Les incendies se produisent généralement en milieu de journée, entre 10 et 18 heures, avec une intensification autour de midi à 16 heures. Cela s'explique par divers facteurs tels que des températures élevées en cours de journée, une activité humaine plus importante (comme des brûlages déclenchés volontairement ou accidentellement) ou des conditions climatiques spécifiques (vent fort, humidité faible) qui favorisent la propagation des incendies pendant ces plages horaires.

Il est constaté qu'il y a très peu d'incendies signalés tôt le matin ou en soirée, ce qui pourrait être attribué à des températures plus fraîches, une humidité accrue ou éventuellement une activité réduite.

Les zones les plus sombres semblent surtout se situer entre 15 et 25 degrés, et entre midi et seize heures, ce qui indique que les conditions dans ce créneau horaire et cette plage de température sont particulièrement défavorables pour la survenue d'incendies.

Les points les plus distincts sont davantage dispersés, mais on en observe plus hors de ces plages critiques de température et d'heure.

Il est observé qu'il existe une corrélation entre la température maximale et le moment de la journée dans l'apparition des feux. La plupart des incendies ont lieu lorsque la température est modérément élevée, entre 15 et 25 degrés, en milieu de journée.

Cela indique que l'accroissement de ces deux éléments accroît le danger des incendies.

L'éparpillement des points indique une variation dans le nombre d'incendies, même au sein des créneaux horaires et des plages de température les plus critiques.

Température : Il a été confirmé que la fourchette de 15 °C à 25 °C est celle qui favorise le plus les incendies, probablement à cause de conditions sèches et chaudes qui stimulent l'inflammation de la végétation.

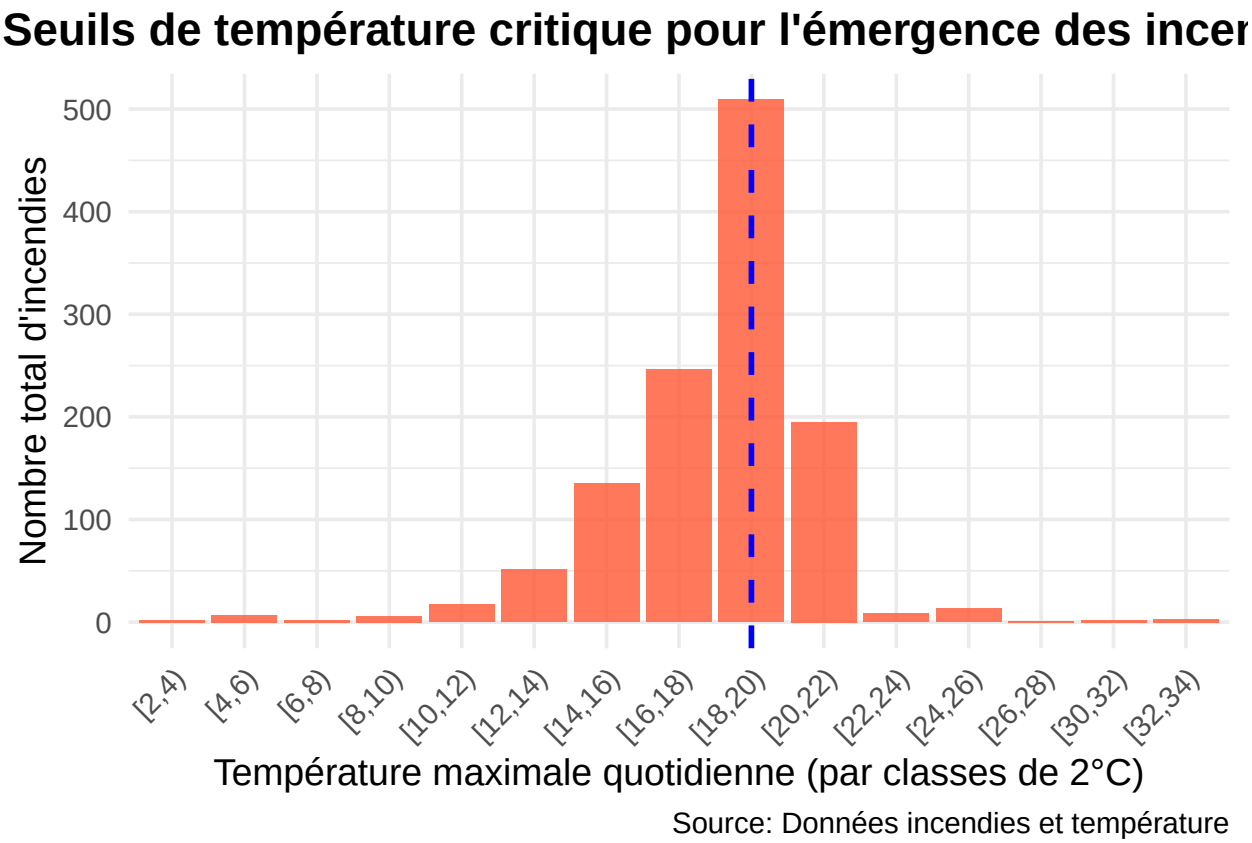
Le pic des incendies entre midi et seize heures est associé aux facteurs suivants :

- Les températures les plus élevées sont fréquemment enregistrées en début d'après-midi, ce qui contribue à la dessiccation de la végétation.
- L'activité humaine (comme les feux de camp, les barbecues ou les mégots de cigarettes mal éteints) est plus courante à ces moments-là.
- Les conditions climatiques (comme le vent ou une humidité faible) peuvent favoriser la diffusion des feux en plein jour.

Les régions affichant le plus grand nombre d'incendies (indiquées par des points rouges foncés) sont celles où les conditions sont les plus critiques, ce qui pourrait orienter les actions de prévention (par exemple, imposer des interdictions de feux de plein air entre midi et 16h lorsque la température excède 15 °C).

3- Seuils de température critique pour l'émergence des incendies:

```r
library(ggplot2)
library(dplyr)

# Charger les données
agg_data <- read.csv("../Exports/export_incendiestempheure.csv")

# Créer des classes de température (par ex. tous les 2°C)
agg_data_bins <- agg_data %>%
  mutate(
    tmax_bin = cut(tmax_med, breaks = seq(0, max(tmax_med, na.rm = TRUE) + 2, by = 2), right = FALSE)
  ) %>%
  group_by(tmax_bin) %>%
  summarise(nb_incendies = n())

# Visualiser : Histogramme des incendies par classe de température
ggplot(agg_data_bins, aes(x = tmax_bin, y = nb_incendies)) +
  geom_col(fill = "#FF5733", alpha = 0.8) +
  geom_vline(xintercept = which(agg_data_bins$nb_incendies == max(agg_data_bins$nb_incendies)),
             linetype = "dashed", color = "blue", size = 1) +
  labs(
    title = "Seuils de température critique pour l'émergence des incendies",
    x = "Température maximale quotidienne (par classes de 2°C)",
    y = "Nombre total d'incendies",
    caption = "Source: Données incendies et température"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



Analyse Informatique

D'abord, on importe les deux bibliothèques indispensables : **dplyr** pour le traitement des données et l'agrégation, et **ggplot2** pour la représentation graphique de ces dernières.

Par la suite, les données sont importées dans une dataframe.

Nous passons ensuite à l'étape de création de classes de température en utilisant la méthode **cut()** qui nous permet de segmenter la variable **tmax_med** en intervalles de 2 degrés, comme par exemple entre 0 et 2 degrés, entre 2 et 4 degrés.

En outre, **seq(0 , max())** nous aidera à établir les limites des classes, allant de 0 jusqu'à la température maximale enregistrée augmentée de 2 degrés.

Ensuite, nous regroupons les données par catégorie de température en employant la méthode **group_by(tmax_bin)**. Par la suite, nous employons **summarise(nb_incendies=n())** afin de déterminer le nombre total d'incendies pour chaque catégorie de température.

Une fois les données prêtes, nous devons passer à la phase de visualisation de l'histogramme des incendies en fonction des classes de température.

Pour ce faire, nous utilisons la méthode **ggplot()**.

Nous construisons un histogramme à barres verticales où chaque barre illustre le nombre d'incendies pour une classe de température donnée en utilisant la technique **geom_col()**.

On choisit l'option de **remplissage** pour assembler les barres.

On insère une ligne verticale en pointillés bleus dénotant la classe ou le nombre d'incendies est au maximum, en employant la fonction prédéfinie **geom_vline()**.

Ensuite, nous utilisons la méthode prédéfinie **labs()**, où nous ajoutons un titre, les libellés des axes et une référence pour le graphique.

On utilise **theme_minimal()** pour donner un aspect sobre et professionnel, et l'on incline les labels de l'axe des X à 45 degrés pour améliorer la clarté en mettant en œuvre la méthode **element_text()**.

Analyse Statistique

Ce graphique vise à nous aider à analyser et à illustrer la distribution des températures maximales journalières, exprimées en degrés Celsius, liées à des incendies.

L'axe des abscisses illustre les températures maximales journalières réparties en catégories de 2 degrés.

L'axe Y indique le total des incendies recensés pour chaque catégorie de température.

Une ligne en pointillés bleus est dessinée autour de 20 degrés, probablement pour indiquer un seuil critique de température.

La distribution présente une **unicité du mode** et est fortement décalée vers la droite. Le maximum est observé dans la tranche **18-20** avec approximativement 500 incendies, ce qui suggère que c'est à cette température qu'un nombre optimal d'incendies se déclenche.

Sur la gauche du sommet, le nombre d'incendies commence à croître graduellement dès que l'on atteint 12 degrés. Après le pic, on observe une chute rapide du nombre d'incendies dès 20 degrés, avec très peu d'incendies au-delà de 28 degrés.

Notre gamme de températures va de 2 à 34 degrés. Toutefois, la plupart de nos incendies se produisent lorsque la température est comprise entre 12 et 26 degrés.

Il est possible d'affirmer que 80% des incendies se déclenchent lorsque la température se situe entre 16 et 22 degrés, atteignant un paroxysme à 18-20 degrés. Les classes 16-18 et 20-22 comptent aussi un nombre élevé de sinistres, approximativement 300 et 200 respectivement.

La ligne en pointillés à 20 degrés indique une limite critique pour le déclenchement des incendies :

- Jusqu'à 20 degrés, le nombre d'incendies s'accroît avec la hausse de la température, signalant une relation positive entre la température et le danger d'incendie jusqu'à ce point critique.
- Au-delà de 20 degrés, la fréquence des incendies chute rapidement, indiquant que des températures plus élevées ne sont pas forcément liées à un risque accru d'incendie.

On pourrait affirmer que 20 degrés représente un seuil critique, où les incendies sont particulièrement courants juste avant et à ce point, mais tendent à diminuer au-delà.

Il est probable que la **moyenne** se situe autour de 18-20 degrés, car c'est là que le pic de la distribution est localisé.

La **mediane** représente la valeur qui sépare les données en deux segments égaux. La température est proche de 18 degrés.

La majorité des données se situe entre 16 et 22 degrés, comme l'indique l'**écart-type**. L'écart type est sans doute entre 2 et 3 degrés.

La **répartition** présente une asymétrie vers la droite, caractéristique des données de températures liées à des phénomènes naturels tels que les incendies.

Nous allons procéder à une interprétation en rapport avec les incendies. Les températures aux alentours de 18-20 degrés semblent être les plus déterminantes pour la naissance des feux.

Ceci peut être dû à un ensemble de facteurs tels que la chaleur adéquate pour assécher la végétation, mais également trop intense pour contrarier des conditions indispensables comme une humidité trop basse ou des vents particuliers.

Au-delà de 20 degrés, le danger diminue dans certains scénarios où la végétation devient trop sèche pour une combustion efficace ou lorsque d'autres facteurs météorologiques, comme des précipitations accompagnées de températures plus élevées, atténuent le risque.

Pour conclure sur notre question générale, nous pourrions affirmer que les incendies surviennent plus souvent lorsque la température est comprise entre 15 et 25 degrés, notamment durant l'heure de la journée où la chaleur est la plus forte, particulièrement entre 10h et 18h.

Cette fourchette de température encourage la déshydratation de la flore et crée des conditions favorables à l'expansion des incendies.

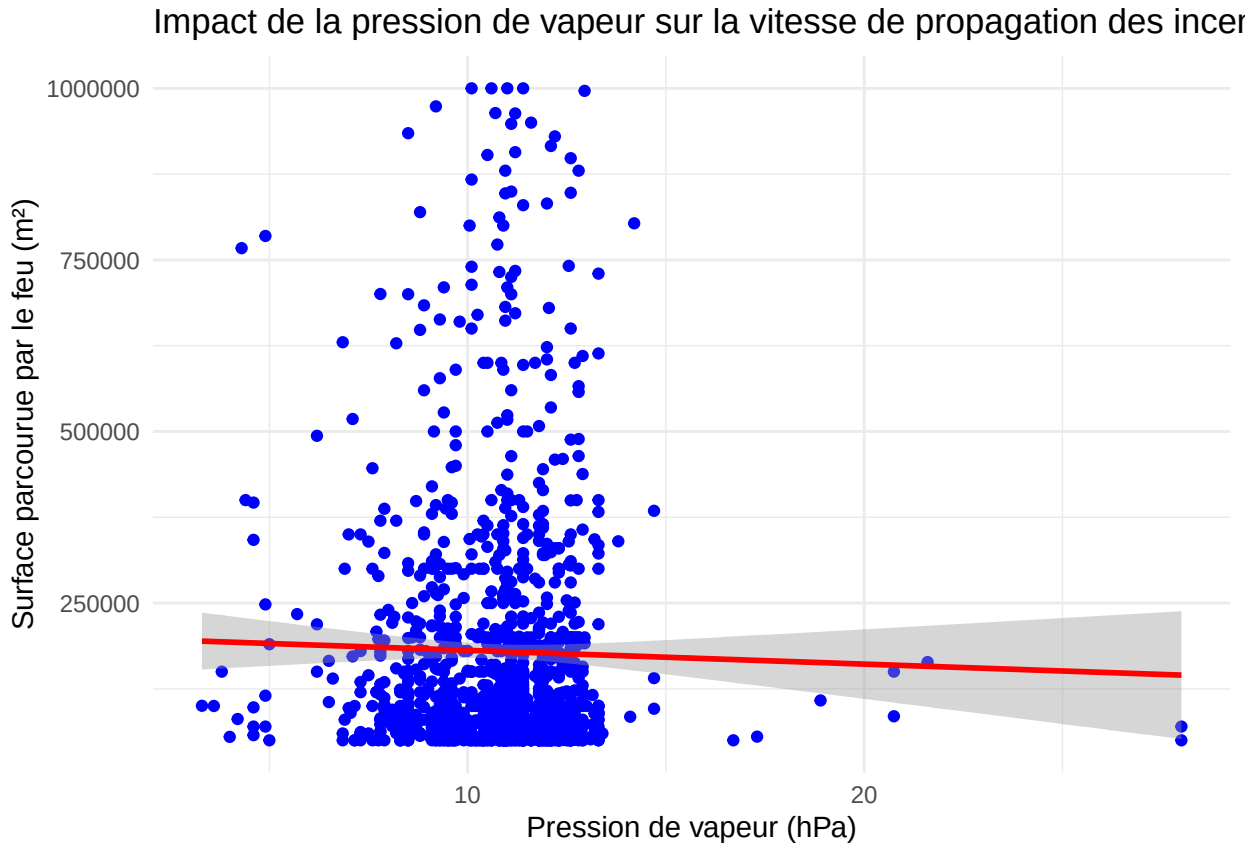
Les incendies sont moins probables à des températures très basses ou très élevées, ce qui souligne l'importance d'une surveillance renforcée et de mesures préventives pendant les heures les plus chaudes de la journée.

5.3.2.7 Relation entre la force du vent et la propagation des incendies

5.3.2.8 Impact de la pression de vapeur sur la vitesse de propagation des incendies Dans cette problématique nous allons se concentrer sur l'impact de la pression de vapeur sur la vitesse de propagation des incendies

```
data <- read.csv("../Exports/export_impactvapeure.csv")

library(ggplot2)

# Créer un graphique de dispersion entre la pression de vapeur et la surface parcourue par le feu
ggplot(data, aes(x = tens_vap_med, y = surface_parcourue_m2)) +
  geom_point(color = "blue") +  # Points bleus
  geom_smooth(method = "lm", color = "red") +  # Ajouter une droite de régression linéaire
  labs(title = "Impact de la pression de vapeur sur la vitesse de propagation des incendies",
       x = "Pression de vapeur (hPa)",
       y = "Surface parcourue par le feu (m²)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



1- **Analyse Informatique:**

Dans un premier temps, nous avons importé la librairie **ggplot2** pour réaliser notre diagramme, puis nous avons intégré nos données à l'aide d'un DataFrame et de la méthode prédéfinie **read.Csv**.

L'objet de données renferme toutes les informations extraites du fichier CSV.

Nous générons le graphique grâce à la librairie **ggplot**, en précisant le jeu de données ainsi que les variables qui seront représentées sur les axes X et Y. De plus, nous définissons les axes via le paramètre **aes**.

On utilise la méthode prédéfinie **geom_point()** pour ajouter les points sur le diagramme de dispersion. Chaque point va symboliser une observation de données.

Nous intégrons une courbe de tendance dans le graphique en employant le paramètre **method="lm"** pour signaler que la courbe représentera une régression linéaire.

Et finalement, en matière de personnalisation, nous attribuons des titres aux deux axes ainsi qu'au titre principal en utilisant **labs()**. Pour lui donner un aspect plus contemporain, nous employons **theme_minimal**.

2- **Analyse Statistique:**

Ce graphique que nous avons est un type de diagramme de dispersion, ou plus précisément, un nuage de points (scatter plot), qui comprend une courbe de régression et une zone d'intervalle de confiance.

L'axe des abscisses indique la pression de vapeur, mesurée en hPa, variant de 0 à 25 hPa, tandis que l'axe des ordonnées représente la vitesse de propagation du feu, exprimée en mètres par seconde sur une échelle logarithmique allant de 0 à 1 000 000 m/s.

Il est à noter que les points de couleur bleue correspondent aux observations individuelles. Dans notre graphique, chaque point illustre une mesure de régression de vapeur et de vitesse de propagation.

Une tendance générale est illustrée par une ligne rouge continue, entourée d'une zone ombragée représentant l'intervalle de confiance supposé à 95%.

Les points montrent une grande dispersion, surtout à basse pression de vapeur entre 0 et 15 hPa, ce qui témoigne d'une variabilité significative dans la vitesse de propagation pour ces niveaux de pression.

Lorsque la pression de vapeur excède 15 hPa, les points se multiplient et la vitesse de diffusion semble généralement se réduire.

La plupart des points se regroupent à des vitesses de propagation inférieures à 250 000 m/s, avec quelques valeurs hors normes atteignant les 1 000 000 m/s.

L'échelle de l'axe des ordonnées est logarithmique, ce qui est courant pour des données très dispersées ou présentant une distribution déséquilibrée, comme c'est le cas ici. Cela nous aide à mettre en évidence les valeurs les plus faibles tout en intégrant les valeurs extrêmes.

Il est possible que la vitesse de propagation des incendies suive une distribution exponentielle ou même log-normale.

Le graphique de régression illustre un lien inverse entre la pression de vapeur et la vitesse de propagation. Quand la pression sur la valeur s'accroît, la vitesse de propagation tend à se réduire.

La courbe présente une inclinaison marquée initialement entre 0 et 10 hPa, puis elle se nivèle progressivement à mesure que la pression de vapeur augmente, entraînant un effet asymptotique où la vitesse de progression atteint une valeur faible pour des pressions élevées.

Pour ce qui est de l'intervalle de confiance, la zone en gris sur le graphique de régression indique cet intervalle.

L'écart se réduit à mesure que la pression de vapeur augmente, car il y a une diminution des données et la variabilité paraît moins importante.

La configuration de la courbe indique un modèle non linéaire, en d'autres termes, une régression logarithmique ou exponentielle, ce qui est en accord avec l'échelle logarithmique de l'axe Y.

Il est hypothétique qu'une transformation logarithmique de la variable dépendante, la vitesse de propagation, a été effectuée pour linéariser la relation, suivie de l'ajout d'un modèle de régression.

On observe une relation inverse entre la pression de vapeur et la vitesse de propagation des incendies. Autrement dit, une pression de vapeur plus élevée, généralement associée à une humidité accrue, tend à freiner la progression des incendies.

En termes physiques, une pression de vapeur plus élevée signale un volume d'eau accru dans l'air, ce qui pourrait diminuer la combustibilité des matériaux et freiner la progression du feu.

L'importante dispersion des points, surtout à basse pression de vapeurs, nous indique que d'autres paramètres non considérés dans ce graphique, tels que la température et le vent, ont un impact sur la vitesse de diffusion.

La courbe de régression indique une tendance décroissante statistiquement significative, puisque son intervalle de confiance ne traverse pas la ligne horizontale nulle. Cela suggère que l'impact de la pression de vapeur sur la vitesse de propagation est effectif et non fortuit.

Pour finir, ce graphique met en lumière une relation inverse manifeste entre la pression de vapeur et la vitesse de propagation des incendies : plus la pression de vapeur est forte, plus la progression des incendies est lente.

5.3.3 Géographie et environnement

```
import geopandas as gpd
import pandas as pd
import matplotlib.pyplot as plt
```

```
df = pd.read_csv("../Data/donnees_geo.csv")

chemin_fichier_geojson = "../Data/contour-des-departements.geojson"
gdf_departements = gpd.read_file(chemin_fichier_geojson)

fig, ax = plt.subplots(figsize=(15, 15))

gdf_departements.plot(ax=ax, color="lightgray", edgecolor="black")

ax.scatter(df["longitude"], df["latitude"], color="red", s=50, label="Incendies")


ax.set_xlim([ -5.5, 10 ])
```

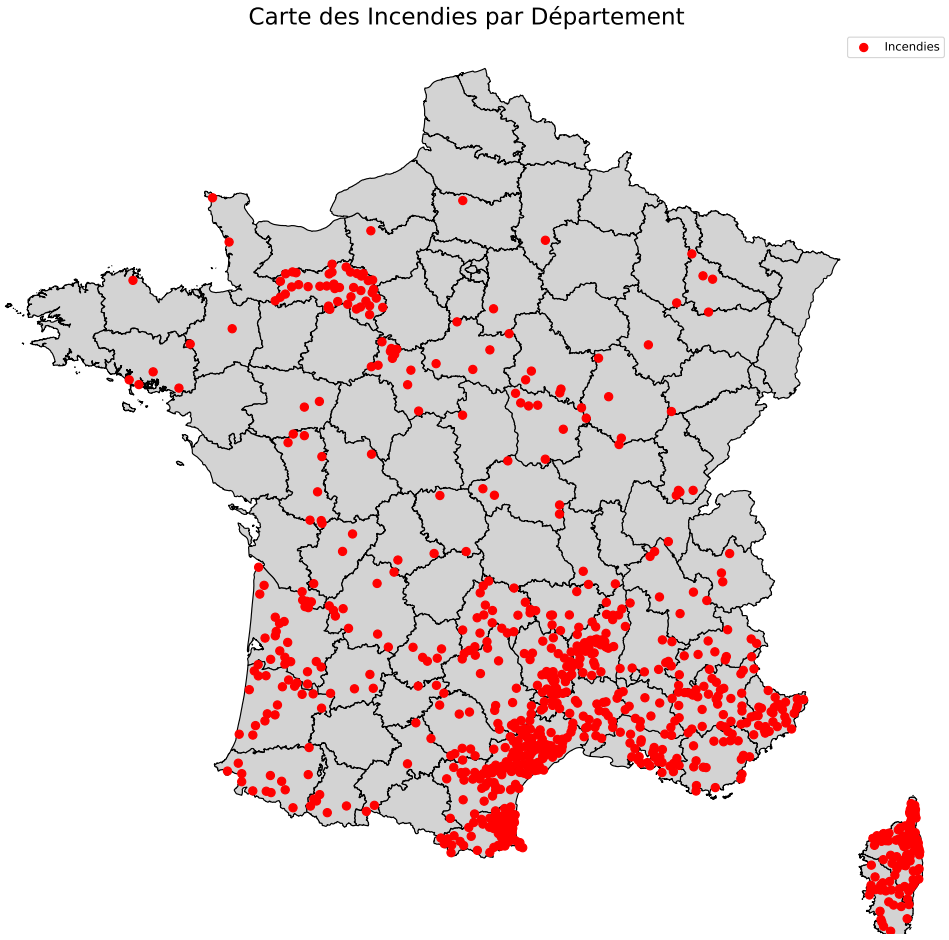
5.3.3.1 Propagation des incendies sur le territoire Francais

```
## (-5.5, 10.0)
```

```
ax.set_ylim([41.5, 51.5])
```

```
## (41.5, 51.5)
```

```
plt.title("Carte des Incendies par Département", fontsize=20)
plt.legend()

ax.set_axis_off()

plt.savefig("carte_incendies_departements_geojson_grande_avec_points_agrandis.png", dpi=600, bbox_inches

plt.show()
```



Nous allons examiner l'analyse de la propagation des incendies à travers le territoire français après avoir utilisé ces informations sur la carte nationale de France.

Décomposons la France, qui est reconnue comme un État composé de régions. Ces régions sont elles-mêmes constituées de sous-régions, qui à leur tour regroupent des départements, et ces derniers, des communes.

En France, on compte 13 régions spécifiquement dans la partie métropolitaine. Nous allons d'abord les nommer, puis commencer leur énumération et enfin procéder à leur analyse.

1. Auvergne-Rhône-Alpes
2. Bourgogne-Franche-Comté
3. Bretagne
4. Centre-Val de Loire
5. Corse
6. Grand Est

7. Hauts-de-France
8. Île-de-France
9. Normandie
10. Nouvelle-Aquitaine
11. Occitanie
12. Pays de la Loire
13. Provence-Alpes-Côte d'Azur

Nous procéderons à l'examen de ces 13 régions individuellement, pour finalement en arriver à une conclusion.

La région **Auvergne Rhône-Alpes** présente un taux d'incendies relativement bas, l'une des raisons étant qu'elle jouit d'un climat plus diversifié que celui d'autres régions du sud.Cette région a des altitudes plus élevées qui entraînent une température plus fraîche et un climat plus humide, ce qui diminue le danger d'incendies de grande envergure.

Le climat tempéré de la région **Bourgogne Franche Comté**, caractérisé par des hivers rigoureux et des étés modérés, explique son taux d'incendies inférieur comparativement à la région Auvergne Rhône-Alpes. Cette région abrite des montagnes et des collines comme le Jura et les Vosges, qui sont habituellement plus humides et moins prédisposées aux incendies forestiers. De plus, cette zone est surtout composée de forêts de feuillus (chênes, hêtres, érables) et de prairies qui présentent une moindre inflammabilité par rapport aux résineux ou aux maquis méditerranéens.

La région de **Bretagne**, en raison de son climat océanique et de ses températures modérées tout au long de l'année, présente un taux réduit d'incendies.Les hivers sont tempérés et les étés restent frais, accompagnés d'une humidité importante. En outre, les précipitations régulières et l'humidité persistante durant toute l'année préservent la végétation plus verdoyante et moins aride, contribuant ainsi à réduire le danger d'incendies.

Il y a effectivement certains incendies dans la région **Centre-Val de Loire**.Cette région jouit d'un climat continental, caractérisé par des hivers rigoureux et des étés plus tempérés en comparaison avec les régions du sud de la France. La région se compose de plaines et de collines, ce qui réduit la probabilité d'incidents majeurs et fait baisser le taux d'incendies dans le centre du Val de Loire.

La Corse est une des zones sensibles aux incendies en raison de divers éléments géographiques, climatiques, etc.La Corse jouit d'un climat méditerranéen marqué par des étés extrêmement chauds et arides, ce qui favorise l'apparition aisée d'incendies.Les températures estivales peuvent atteindre des sommets extrêmes, créant un environnement propice à la diffusion rapide des incendies. Cette zone peut être exposée à des rafales puissantes comme le Mistral qui facilitent la diffusion des feux.Ces courants d'air puissants et généralement secs peuvent propulser les flammes sur de vastes étendues, compliquant ainsi l'éradication des incendies.

La région **Grand-est** présente une moindre susceptibilité aux incidents, mais elle n'est pas totalement épargnée.Cette région jouit d'un climat diversifié qui subit des influences à la fois océaniques et continentales.Les hivers sont rigoureux, les étés sont torrides, cependant ils ne parviennent pas à égaler ceux du sud de la France en termes de températures.L'incidence des périodes de sécheresse est en baisse, diminuant ainsi le danger des incendies. En outre, cette région est dotée de magnifiques zones montagneuses et de vastes plaines. Cette région bénéficie d'une pluviométrie plutôt stable, surtout en automne et en février. Ces précipitations contribuent à conserver un certain degré d'humidité dans le sol et la végétation, ce qui diminue les risques d'incendies.

Le climat humide, la présence de forêts de feuillus, le relief plat et les vents modérés contribuent à rendre la région des **Hauts-de-France** relativement moins vulnérable aux incendies.Ainsi, les conditions météorologiques et géographiques sont généralement moins favorables aux incendies sévères que d'autres zones situées au sud.

L'Île-de-France, étant la région où se trouve la **capitale**, est la plus densément peuplée et urbanisée de France. Elle est moins sujette aux grands incendies que les régions situées au sud du pays. En revanche, cette région n'est pas entièrement protégée contre les incendies. Cette région jouit d'un climat océanique tempéré caractérisé par des hivers rigoureux et des étés doux. En outre, la topographie de l'Ile-de-France

est principalement plane ou légèrement vallonnée, en contraste avec les régions montagneuses telles que les Alpes ou les Pyrénées.

La région de **Normandie** affiche un taux réduit d'incendies.Grace à son climat océanique tempéré, son niveau de précipitations constant, sa topographie relativement plane et sa végétation essentiellement feuillue, la Normandie est moins sujette aux incendies que d'autres zones en France.Ces facteurs contribuent à une augmentation de l'humidité dans l'environnement, réduisant par conséquent le risque d'incendies. Néanmoins, des incendies sporadiques peuvent encore survenir, principalement dans les régions boisées et rurales pendant les périodes de sécheresse ou de chaleur intense, même s'ils restent relativement peu fréquents par rapport aux zones plus chaudes et sèches du pays.

La région **Nouvelle Aquitaine**, située dans le sud-ouest, est susceptible de subir des incendies, notamment à cause de son climat et de sa végétation. Cette région présente un climat plutôt océanique à l'ouest et méditerranéen à l'est, ce qui signifie que les étés peuvent être particulièrement chauds et secs, surtout dans les zones limitrophes de l'Espagne et du Pays Basque. Par ailleurs, les vagues de chaleur peuvent entraîner des phases d'aridité prolongées qui créent un environnement propice aux incendies.

L'Occitanie est une des régions les plus vulnérables aux incendies en raison de divers facteurs géographiques, climatiques et écologiques.Dans sa partie méridionale et sud, cette région jouit d'un climat méditerranéen avec des étés chauds et secs.Durant l'été, les températures peuvent grimper à des sommets élevés et la région est susceptible de connaître des phases d'aridité prolongées, des circonstances propices aux incendies. Dans la propagation des incendies, le vent, notamment le Mistral et la Tramontane qui balaye occasionnellement la zone, a un impact crucial.

La région du **Pays de la Loire** présente une exposition moindre aux incendies par rapport à d'autres régions. Le climat de cette zone est océanique, c'est-à-dire qu'il offre des hivers doux et des étés tempérés, avec des précipitations régulières tout au long de l'année. Les vagues de chaleur se produisent généralement dans d'autres zones, comme le sud-est de la France, ce qui rend la région moins susceptible aux incendies liés à la canicule.

Pour conclure, la dernière région est la **Provence Alpes Côte d'Azur**, qui se trouve dans le sud-est et subit une forte influence de divers facteurs climatiques et géographiques.Cette région jouit d'un climat méditerranéen, marqué par des étés très chauds et secs ainsi que des hivers doux et humides. Durant l'été, les températures peuvent atteindre des niveaux extrêmement élevés avec des vagues de chaleur qui se produisent fréquemment. On peut également observer que cette zone est celle où le taux d'incendies est le plus élevé.

Suite à l'étude de la progression du taux d'incendies sur la carte nationale que nous avons réalisée, il est évident que la région où le taux d'incendies est le plus élevé est la **Côte d'Azur**. En revanche, les zones où le taux d'incendies est le plus faible incluent des régions telles que les **Pays de la Loire**, **Normandie**, etc. Il est donc clair que les conditions météorologiques et la localisation géographique de la ville ou de la région ont un impact crucial, surtout lorsque les conditions climatiques sont particulièrement propices aux incendies. De même, la nature des forêts situées dans les zones urbaines a une incidence significative sur l'expansion des feux.Bien qu'il existe une différence de taux d'incendies entre la ville et la montagne, cette question sera examinée dans le cadre d'une autre problèmatique.

5.3.3.2 Comparaison de la fréquence des incendies dans les régions montagneuses vs basses altitudes L'objectif de cette étude est d'examiner et de comparer le lien entre les incendies et l'altitude dans diverses régions et départements en France.

Les données utilisées sont issues de deux sources majeures. Un fichier CSV contient divers attributs, parmi lesquels nous avons choisi ceux qui nous intéressent, à savoir les données concernant la superficie des terrains brûlés, la nature de ces feux, ainsi qu'un fichier géographique comprenant des renseignements sur l'altitude moyenne de chaque municipalité.

Cette analyse vise à déterminer si les incidents ont tendance à être plus fréquents ou plus graves aux altitudes élevées comparativement aux altitudes inférieures.

Avant de procéder à l'analyse, nous allons décrire nos données.

Les informations sur les incendies incluent des données concernant la superficie des incendies exprimées en **surface_parcourue_m2**, ainsi que le type d'incendie principal et secondaire indiqué soit par **nature_inc_prim**, soit par **nature_inc_sec**. De plus, nous disposons de détails chronologiques tels que **année, mois, jour, heure**.

Le fichier **donnes_geo** nous donne des renseignements sur la hauteur moyenne de chaque commune, illustrée en **altitude_med**, ce qui nous aidera à catégoriser les incendies selon leur altitude en deux options : soit **haute**, soit **basse** altitude.

Dans cette analyse cruciale, nous avons utilisé une variable catégorielle **altitude_zone** qui classe chaque commune selon son altitude moyenne : soit en zones de **basse altitude** en utilisant une délimitation de **moins de 1000 mètres** que nous avons employée pour les altitudes inférieures, soit en zones de **haute altitude** grâce à une dénomination **au-dessus de 1000 mètres.**

Apres avoir decrits touts els donnees qu'on possede pour cette anaylyse on va commencer a calculer la **correlation de Pearson**. On va calculer la correlation entre la surface des incendies et l'altitude. Le but de cella la est de voir si il existe **une relation lineaire** entre la surface des incendies et l'altitude.

On va y calculer la correlation et apres on va analyser le graphe de la correlation de Pearson

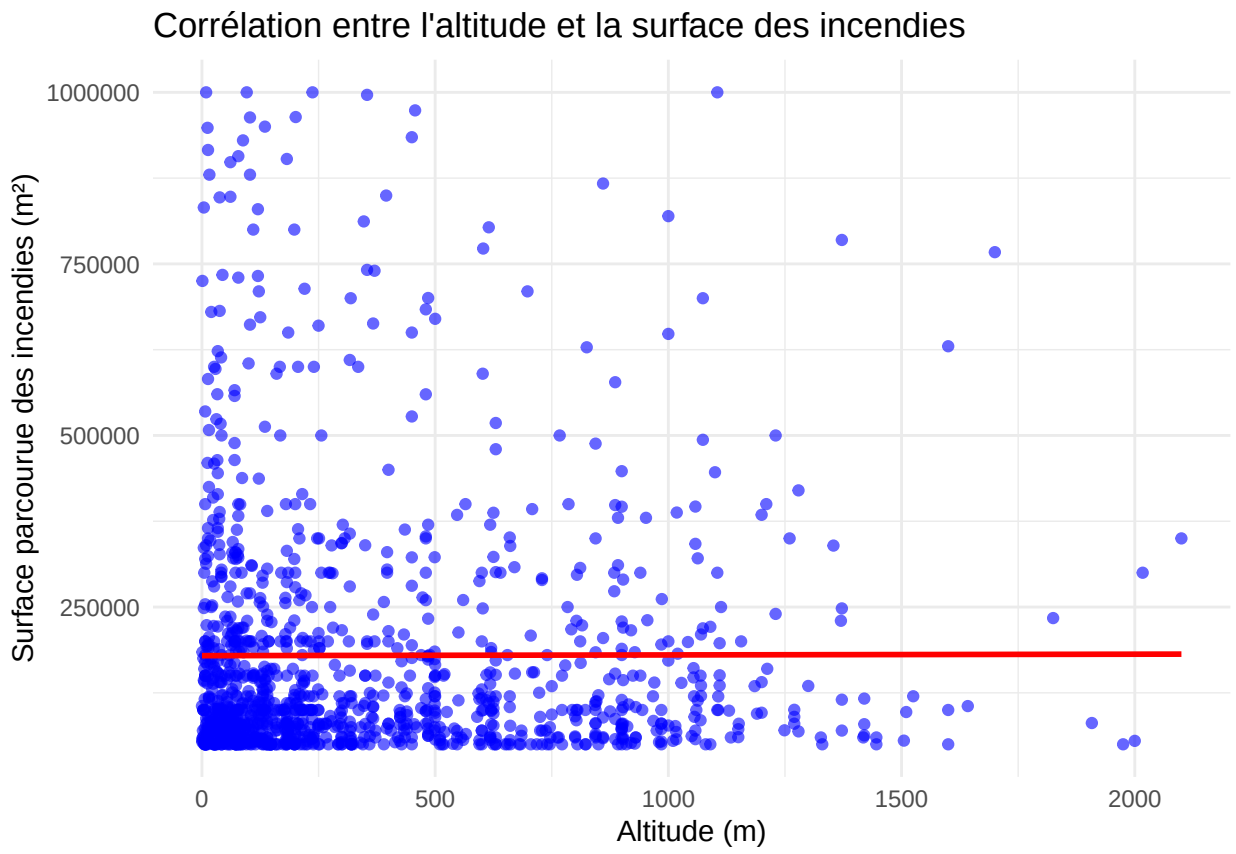
```
incendiesregions <- read.csv("../Exports/export_incendiesregions.csv")
correlation <- cor(incendiesregions$surface_parcourue_m2, incendiesregions$altitude_med, use = "complete

print(paste("Corrélation de Pearson entre la surface des incendies et l'altitude : ", round(correlation
```

```
## [1] "Corrélation de Pearson entre la surface des incendies et l'altitude :  0"
```

Nous observons une corrélation de 0, ce qui indique qu'il n'y a pas de lien linéaire entre les deux variables. Nous concluons donc qu'elles sont indépendantes l'une de l'autre.

```
incendiesregions <- read.csv("../Exports/export_incendiesregions.csv")
ggplot(incendiesregions, aes(x = altitude_med, y = surface_parcourue_m2)) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(title = "Corrélation entre l'altitude et la surface des incendies",
       x = "Altitude (m)", y = "Surface parcourue des incendies (m²)") +
  theme_minimal() +
  geom_smooth(method = "lm", se = FALSE, color = "red")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



1- **Analyse Informatique du Graphe de Corrélation**

On a commencé par charger le fichier CSV (valeurs séparées par des virgules) en utilisant la méthode prédéfinie **read.csv**. Nous avons également eu recours à la bibliothèque ggplot2 pour réaliser le graphique qui sera ensuite soumis à une analyse statistique.

De plus, nous avons précisé dans la méthode **ggplot** le paramètre contenant le fichier CSV ainsi que l'**aesthetics** d'où nous indiquons les axes X et Y.

Par la suite, on utilise geom-point() pour intégrer des points (**scatter plot**) au diagramme.Dans notre schéma, chaque duo illustre une paire de valeurs.L'option color donne la faculté de changer la couleur des points en bleu, tandis que l'option alpha permet de définir leur degré de transparence.

Un **alpha** de 1 signifierait une opacité totale, alors qu'un alpha de 0 indiquerait une transparence totale.

Par la suite, nous employons la fonction **labs()** pour attribuer des étiquettes au graphique et nous intégrons la méthode **theme_minimal** afin d'appliquer un style contemporain à notre diagramme.

Pour appliquer le graphique de corrélation, nous employons la méthode **geom_smooth()** qui nous donne l'opportunité d'intégrer la ligne de tendance, également connue sous le nom de courbe de régression, sur le graphique.Cela vise à mettre en évidence une tendance globale entre les variables.

Nous aimerions faire un récapitulatif des options disponibles dans la méthode geom_smooth() :

- L'option **method = "lm"** indique qu'une régression linéaire est employée. Cette régression va nous révéler le lien linéaire entre l'altitude et la superficie des feux de forêt.
- L'option **se=FALSE** nous permet d'éviter l'affichage de l'intervalle de confiance.
- L'option **color = "red"** sert à définir la couleur de la courbe de régression en rouge.

2- **Analyse Statistique:**

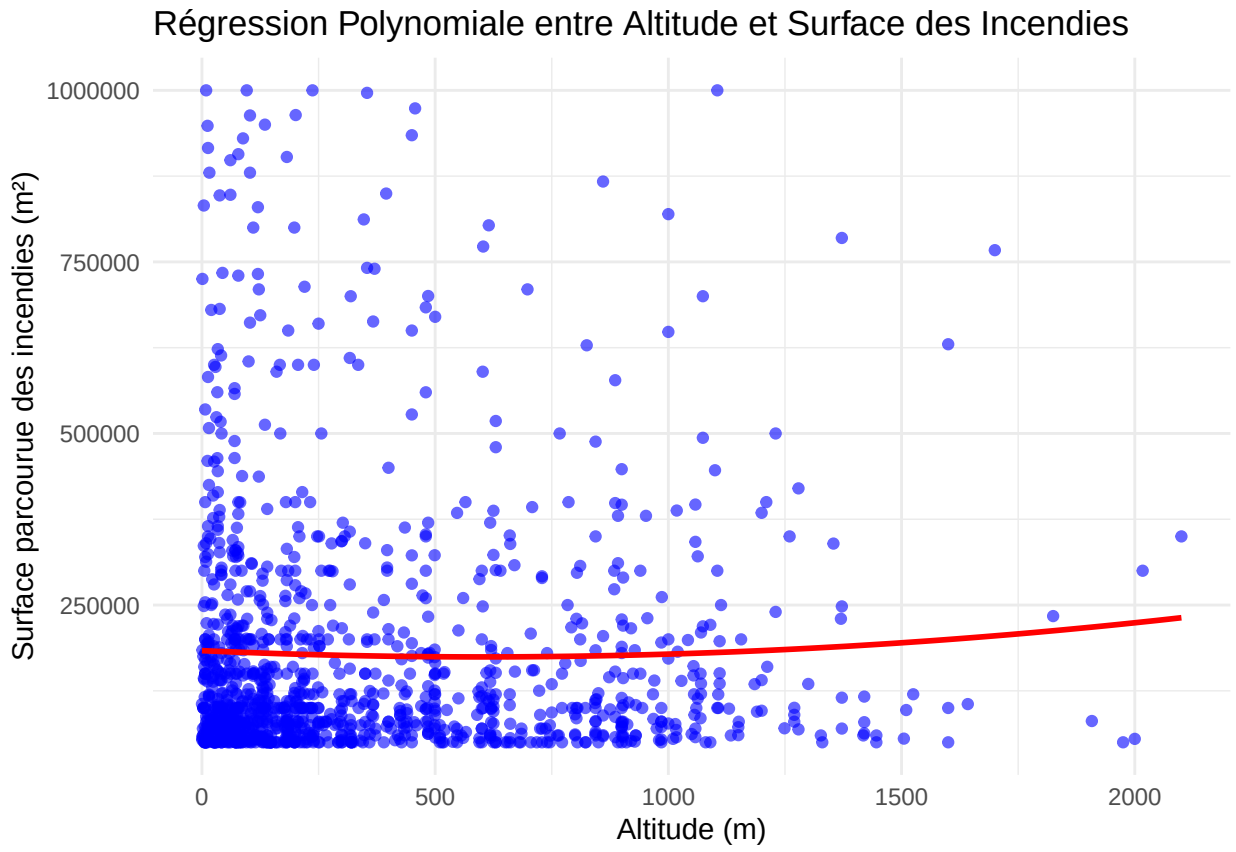
En examinant notre diagramme de façon statistique, nous notons que l'Axe des X illustrant l'altitude s'étend de 0 à 2000 mètres. Concernant l'axe des Y, il indique la superficie des incendies allant de 0 à 1 000 000 m², même si la plupart des points se regroupent sous les 750 000 m². Nous précisons que les points bleus sur notre graphique correspondent aux observations individuelles de la superficie des incendies à diverses altitudes.

On note une importante densité de points à des altitudes faibles (proches du niveau de la mer), accompagnée d'une grande variabilité dans la superficie des incendies, qui peut atteindre environ 750 000 m². Quand l'altitude dépasse 500m, la quantité de points diminue et l'étendue des incendies paraît généralement moins importante avec très peu de cas où les feux excèdent 250,000 m2 au-delà de 1000m. La ligne rouge horizontale, placée autour de 250,000 m², représente une référence ou, pour dire autrement, la moyenne.

On observe une tendance à la baisse de la superficie des incendies lorsque l'altitude augmente. On pourrait également affirmer que le taux d'incendies en montagne sera inférieur à celui observé en milieu urbain. Il est crucial de noter que l'écart entre les points est significatif, surtout à basse altitude, ce qui suggère que le lien n'est pas fortement linéaire. Cela peut être influencé par divers éléments tels que le climat, la végétation, l'activité humaine, etc., qui pourraient affecter l'étendue des incendies.

Une fois la corrélation de Pearson calculée, nous allons maintenant procéder à la mise en œuvre d'une régression polynomiale, à condition que nous admettions que les deux variables sont indépendantes linéairement.

```
incendiesregions <- read.csv("../Exports/export_incendiesregions.csv")
ggplot(incendiesregions, aes(x = altitude_med, y = surface_parcourue_m2)) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(title = "Régression Polynomiale entre Altitude et Surface des Incendies",
       x = "Altitude (m)", y = "Surface parcourue des incendies (m²)") +
  theme_minimal() +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), se = FALSE, color = "red")
```



On va en premier analyser en premier une Analyse Informatique et puis une Analyse Statistique:

1- **Analyse Informatique:**

Nous utilisons également la bibliothèque **ggplot2** ici. Nous commençons à construire notre graphique en invoquant la méthode **ggplot()**, tout en précisant le DataFrame. En définissant l'**aes**, nous déterminons les axes x et y. Ensuite, nous traçons les points sur le graphique en spécifiant leurs couleurs et transparences. Par la suite, nous ajoutons les étiquettes à notre graphique et appliquons le **theme_minimal()** afin de lui conférer une apparence moderne.

On finit par tracer la courbe de la régression polynomiale en employant la méthode **geom_smooth()**. Il est précisé que la méthode employée sera la régression linéaire, puis on note le score de la formule de régression polynomiale en utilisant poly(x,2). On indique que l'on souhaite x et x2, pas seulement x. **Se** implique que nous masquons l'intervalle de confiance.

2- **Analyse Statistique:**

Nous allons examiner le graphique de la régression polynomiale entre l'altitude et la superficie des feux de forêt. L'axe des x va de 0 à 2000 mètres, tandis que l'axe des y s'étend de 0 à 1 000 000 m², avec une grande partie des points situés au-dessus de 750 000 m².

Les observations individuelles sont représentées (la superficie des incendies à diverses altitudes).En outre, la ligne rouge qui entoure environ 250,000 m² représente la régression polynomiale.

La régression des points correspond à l'analyse antérieure du graphique, présentant une altitude basse de 0 à 500m avec une large variabilité allant jusqu'à 750,000 m2. On observe une diminution graduelle de la densité des points et des surfaces à mesure que l'altitude s'accroît.

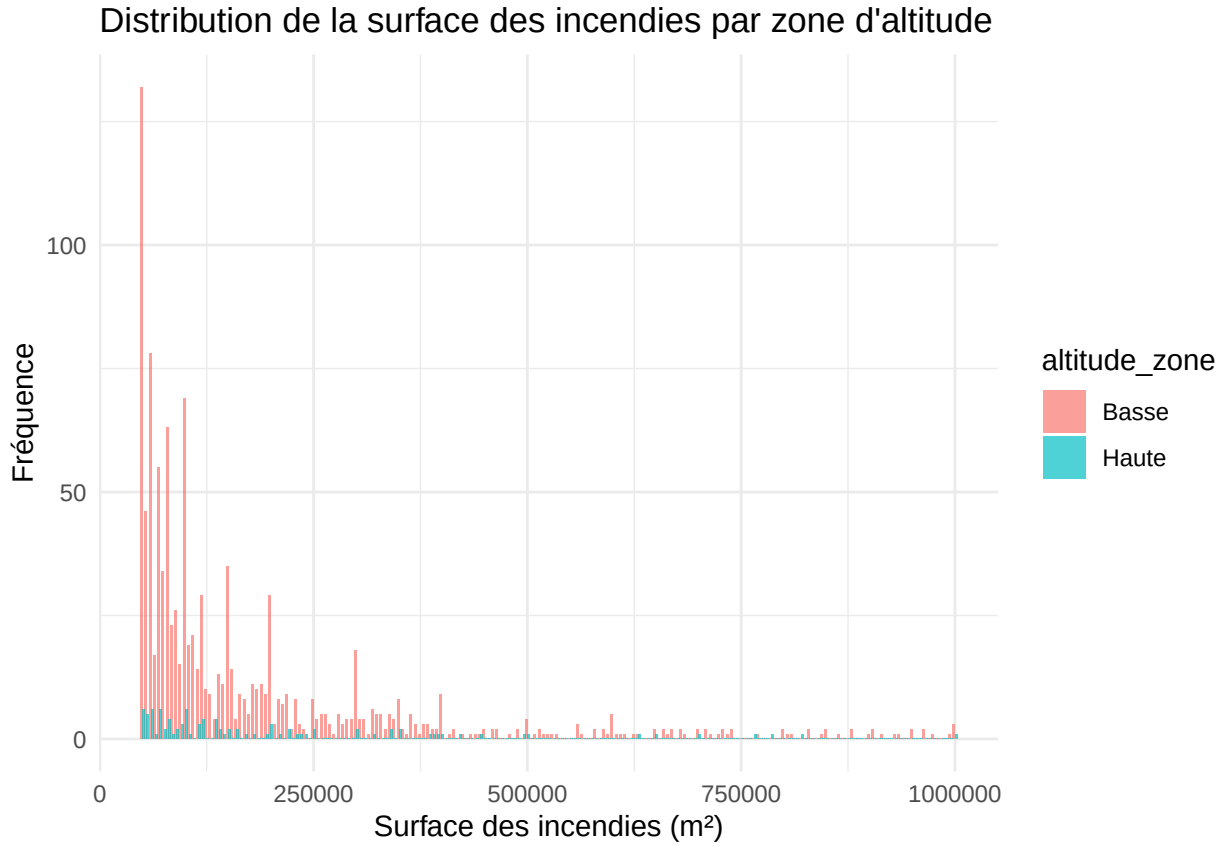
La courbe de régression polynomiale est une ligne légèrement montante.

Après avoir effectué ces deux calculs, nous passerons à l'étape d'analyse approfondie de nos données.Nous commencerons par examiner la surface d'altitude par zone d'altitude.

1. **L'analyse de la surface des incendies par zone d'altitude**

Pour ce faire, nous allons construire notre histogramme puis réaliser deux types d'analyses : une analyse informatique et une autre statistique.

```
incendiesregions <- read.csv("../Exports/export_incendiesregions.csv")
ggplot(incendiesregions, aes(x = surface_parcourue_m2, fill = altitude_zone)) +
  geom_histogram(binwidth = 5000, position = "dodge", alpha = 0.7) +
  labs(title = "Distribution de la surface des incendies par zone d'altitude",
       x = "Surface des incendies (m²)", y = "Fréquence") +
  theme_minimal()
```



1- **Analyse Informatique:**

Pour dessiner cet histogramme, nous avons utilisé la bibliothèque **ggplot2**.Dans un premier temps, nous avons importé les données du fichier CSV dans une variable sous forme de DataFrame.

Par la suite, nous avons utilisé la méthode ggplot().Cette dernière nécessite en paramètres la variable où les données sont entreposées ainsi que l'aes où nous spécifions nos deux axes, x et y, le x représentant ici la surface des incendies. Nous avons ensuite coloré les barres de l'histogramme en fonction des différentes zones d'altitude.

Pour intégrer l'histogramme, nous avons utilisé la méthode **geom_histogram** qui nécessite plusieurs paramètres. Parmi ces paramètres, nous indiquons la largeur des classes ou en d'autres termes, le **binwidth**, qui sera fixé à 5000 m² pour agrégater les incendies en segments de 5000 m². Nous spécifions

également que les barres seront disposées en mode **dodge**, c'est-à-dire côte à côte pour chaque zone afin de faciliter la comparaison entre les différentes catégories.Et finalement, nous précisons que la transparence des couleurs sera de **0,7**, afin de simplifier l'interprétation de notre histogramme.

Ensuite, nous intégrons le titre et les étiquettes à notre histogramme en utilisant la méthode prédéfinie **labs()**. Cette dernière nécessite deux paramètres : title pour attribuer un titre au graphique, ainsi que x et y pour nommer respectivement les axes des abscisses et des ordonnées.

Pour personnaliser le thème, nous avons employé la méthode prédéfinie **theme_minimal()** pour conférer un style contemporain.

2- **Analyse Statistique:**

Les incendies constituent un défi majeur pour la gestion des risques liés à l'environnement, aux personnes et à la sécurité. Comprendre leur distribution spatiale, en particulier par rapport à l'altitude, nous aide à repérer les zones sensibles et à ajuster nos stratégies de prévention.

Cette étude vise à examiner les informations fournies sur les surfaces brouillées dans quatre zones d'altitude, allant de la basse à la haute altitude.

Nous visons à identifier les tendances et à examiner leurs causes potentielles.

La description des données mentionne quatre valeurs principales qui correspondent aux zones brûlées.

La plupart des feux dans ces deux zones se situent sur des surfaces plutôt réduites, allant de 0 à 250 000 m^2. Cela est signalé par les fréquences élevées.

Alors que la superficie s'accroît au-delà de 250,000 m^2, le taux d'incendies chute rapidement pour les deux zones, avec très peu d'entre elles dépassant les 500,000 m^2.

Nous allons maintenant présenter la comparaison entre les deux principales zones, Zone Basse et Zone Haute.

Nous allons débuter avec la Zone Basse. Cette région affiche une occurrence nettement supérieure, en particulier pour les petites surfaces, avec un sommet frôlant les 100.On peut donc supposer que les incendies de petite à moyenne envergure se produisent beaucoup plus souvent en zone basse.En d'autres termes,on peut constater que les incendies surviennent plus souvent dans les villes ou les communes situées à une altitude basse.

Concernant la région à altitude élevée.La survenue de ces incendies est moins fréquente,avec des sommets pouvant aller jusqu'à 10 - 20 sur une même étendue de terrain.

Cela suggère que les incendies sont moins courants dans les zones élevées, ou que leur portée est plus restreinte.

Pour le dire autrement, on peut affirmer que la fréquence des incendies est plus élevée dans les villes ou communes situées à une altitude importante.

Suite à l'étude des deux régions en altitude, nous allons discuter de la tendance globale.

La distribution présente une forme d'atténuation exponentielle que l'on pourrait qualifier de queue lourde selon les statistiques, pour ces deux zones.

On peut également affirmer que la probabilité de voir se produire des incendies de grande ampleur dépassant les 500,000 m2 est extrêmement faible.

Il est à noter que, au-delà de **750,000 m2**, les incendies d'une très grande envergure sont peu fréquents dans ces deux régions.

Le mode en statistique, représentant la valeur la plus courante dans nos données, est positionné sur la première barre pour les deux zones. Cependant, son accentuation est davantage marquée en zone basse.

La variation des surfaces brûlées est plus significative dans les zones où les incendies englobent une vaste étendue de terrain avec une fréquence remarquée. Dans les zones élevées, la dispersion est moindre avec une concentration autour des petites surfaces.

La distribution présente une forte asymétrie à droite (avec une biais positif) et possède une longue queue pour les grandes superficies, même si cette dernière est pratiquement nulle au-delà de 500,000 m2.

Il est probable que les incendies dans les zones basses soient plus courants en raison des conditions environnementales propices (végétation dense, températures élevées, accès humain facilité).

La fréquence des incendies en zones élevées pourrait être due à des éléments tels qu'une végétation moins dense, un climat plus frais ou une accessibilité restreinte.

L'occurrence rare de vastes incendies dépassant les 500,000 m2 dans ces deux régions pourrait témoigner de mesures efficaces ou de barrières naturelles (comme la topographie et l'humidité).

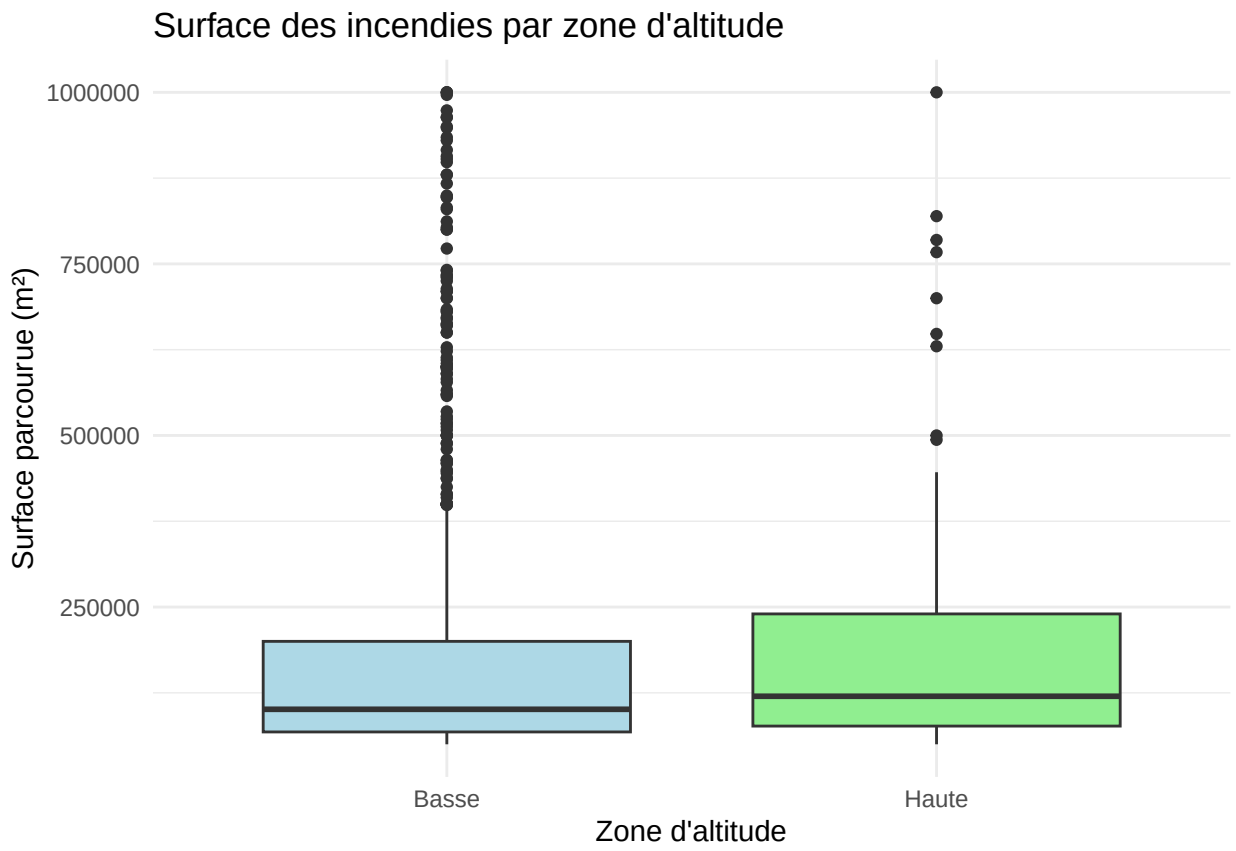
On pourrait finalement convenir que ce graphique indique une fréquence et une variété nettement plus grandes des incendies en termes de superficie dans les zones de basse altitude, avec une prédominance marquée des petits incendies qui se propagent entre 0 et 250,000 m^2.

Dans les zones élevées, les incendies sont moins fréquents et se restreignent généralement aux petites superficies.

2. **La surface des incendies par zone haute et baisse des altitudes**

Donc en premier on va dresser notre **boxplot**. On indique que le boxplot il est utile pour pouvoir visualiser la repartition des donnees.

```
incendiesregions <- read.csv("../Exports/export_incendiesregions.csv")
ggplot(incendiesregions, aes(x = altitude_zone, y = surface_parcourue_m2)) +
  geom_boxplot(fill = c("lightblue", "lightgreen")) +
  labs(title = "Surface des incendies par zone d'altitude",
       x = "Zone d'altitude",
       y = "Surface parcourue (m²)") +
  theme_minimal()
```



1- **Analyse Informatique:**

D'abord, nous soulignons que ce diagramme en boîte, également connu sous le nom de **boxplot** en anglais, nous donnera la possibilité d'observer la répartition des surfaces brûlées dues à des feux dans deux principales zones.

Comme à l'accoutumée, nous importons les données dans une variable sous forme de DataFrame en spécifiant le chemin relatif du fichier contenant ces données.

Par la suite, on utilise la méthode **ggplot** pour créer un graphique initial à l'aide du package **ggplot2**.Ainsi, dans cette approche, on recourt à la variable qui contient les données précédemment stockées comme paramètre.Ensuite, nous définissons l'altitude comme l'axe des x et nous alignons les ordonnées sur l'axe des y.

On précise que **altitude_zone** fait référence aux catégories d'altitudes, qu'elles soient basses ou élevées, et que **surface_parcourue_m2** représente la superficie brûlée en mètres carrés.

Pour intégrer le graphique dans le box, on emploie la méthode préétablie de la bibliothèque **geom_boxplot()**. qui va produire un boxplot, une forme de représentation graphique qui permet d'illustrer la dispersion des données, y compris la valeur médiane, etc. Par ailleurs, l'option fill nous autorisera à spécifier les teintes pour remplir les boutes.

On attribue finalement des titres et des étiquettes aux axes ainsi qu'au graphique en utilisant la fonction prédéfinie **labs()**. Pour personnaliser et moderniser le thème, comme à l'accoutumée, nous utilisons la fonction prédéfinie **theme_minimal**.

2- **Analyse Statistique**

Nous allons procéder à une analyse statistique de notre graphique qui est illustré par un **diagramme en boîte**.Nous allons comparer la répartition de la superficie des incendies exprimée en m² dans deux régions d'altitude : basse et haute.

Nous débutons la présentation des composantes de notre graphique avec l'axe Y, qui illustre la superficie parcourue en m² sur une échelle allant de 0 à 1 000 000 m².L'axe X dépeint quant à lui les deux catégories : Basse et Haute.

Notre graphique présente deux boîtes, chacune d'elles illustrant la répartition des données avec des moustaches et des points qui symbolisent les valeurs extrêmes. Il est à noter que la zone **Basse** est indiquée en **bleu clair** et la zone **Haute** en **vert pâle**.

Nous allons définir un **BoxPlot** comme contenant les informations suivantes :

- La ligne médiane dans la boîte est un indicateur de la **Médiane**.
- Les **Quartiles** sont les extrémités de la boîte qui illustrent le premier quartile.
- Les **Moustaches** s'étendent habituellement jusqu'aux valeurs extrêmes, maximales et minimales.
- Les **Points** symbolisent les valeurs extrêmes qui dépassent les moustaches.

La Zone Basse englobe la région médiatique où elle semble se déployer sur environ 200,000 m2. Les moustaches de cette zone basse s'étendent jusqu'à 50,000 m en profondeur et 500,000 m2 en hauteur.

De plus, on peut observer une multitude de points entre 500,000 et 1,000,000 qui signalent des incendies d'une ampleur exceptionnelle.

Concernant la Zone Haute, sa superficie s'étend approximativement de 100,000 m2 à 250,000 m2.

De plus, elle semble légèrement inférieure à celle de la Zone Basse qui est d'environ 150,000 - 175,000 m2.

Les moustaches se déploient sur une superficie d'environ 50 000 m² en bas et de 500 000 m² en haut.

On constate que moints de points abeerant que dans la Zone Basse avec quelques valeurs entre 500,000 et 750,000 m2.

En observant les deux zones, on remarque :

- On constate que la **mediane** est supérieure dans la **Zone Basse**, avoisinant les 200,000 m2, par rapport à la zone Haute, où elle se situe entre 150,000 et 175,000 m2. Cela indique que les incendies ont tendance à occuper une superficie plus importante en altitude inférieure.
- Les deux boîtes sont asymétriques du côté droit.
- La Zone Basse présente davantage de valeurs extrêmes jusqu'à 1,000,000 m2 comparativement à la Zone Haute qui atteint jusqu'à 750,000 m2. Cela pourrait indiquer que les conditions sont propices à des incendies hors du commun en basse altitude.

Il est possible que les incendies plus importants en basse altitude soient attribuables à une abondance accrue de combustible (forêts denses, broussailles) ou à une facilité d'accès pour des facteurs humains (déclenchements de feu).

À haute altitude, les conditions plus sévères pourraient entraver la propagation.

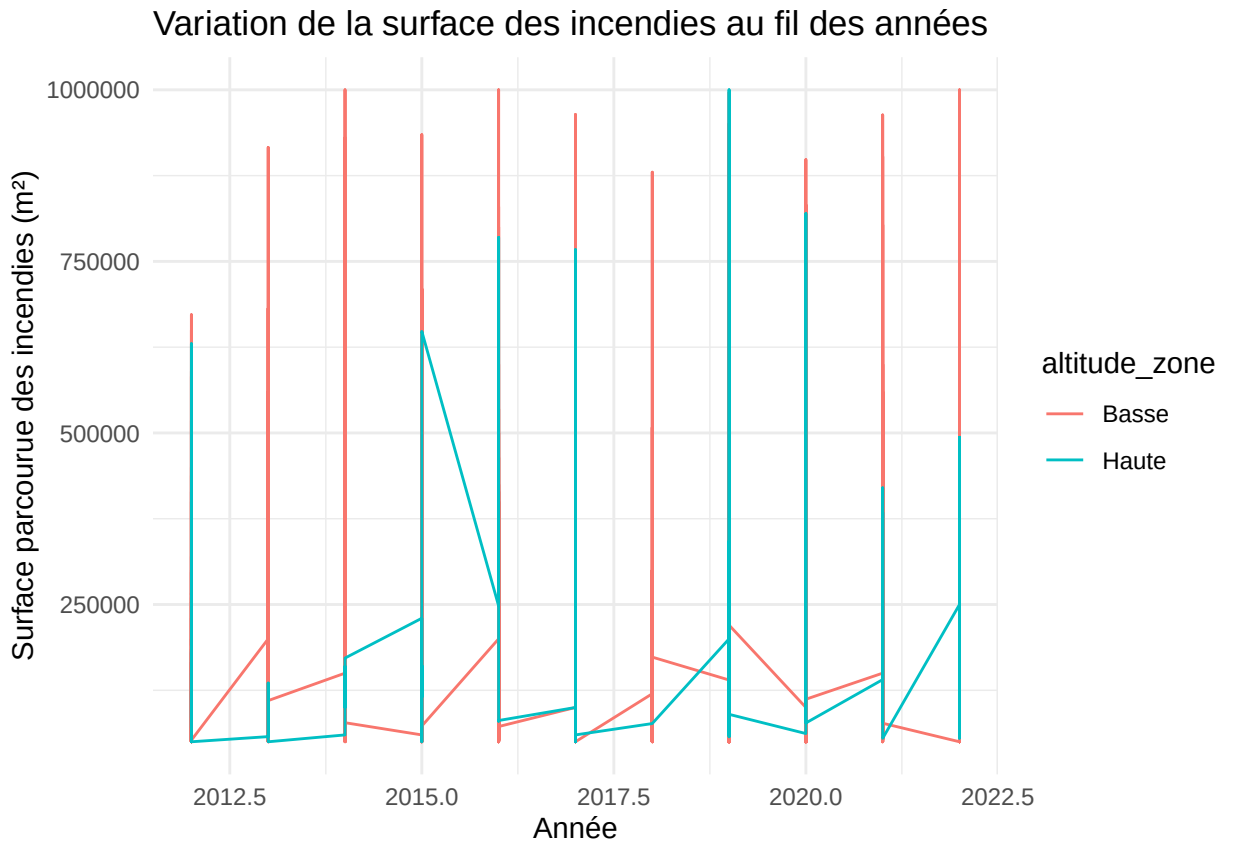
Il est possible de déduire à partir de cette sous-problématique que ce graphique illustre une disparité dans l'étendue des incendies en fonction de l'altitude.

Les régions de basse altitude présentent une superficie parcourue nettement plus grande que les zones situées en haute altitude.

Cette variation est biologiquement justifiée par l'augmentation de la densité de végétation combustibles, une intervention humaine plus prononcée ou même des conditions météorologiques plus favorables au déclenchement d'incendies dans les zones basses.

3- **La variation de la surface des incendies au fils des années:**

```r
incendiesregions <- read.csv("../Exports/export_incendiesregions.csv")
ggplot(incendiesregions, aes(x = annee, y = surface_parcourue_m2, color = altitude_zone)) +
  geom_line() +
  labs(title = "Variation de la surface des incendies au fil des années",
       x = "Année", y = "Surface parcourue des incendies (m²)") +
  theme_minimal()
```



1- **Analyse Informatique:**

Nous commencerons par réaliser une analyse informatique en détaillant la façon dont nous avons programmé ce graphique, puis nous passerons à l'analyse scientifique.

Dans un premier temps, nous avons assigné les données sous forme de tableau DataFrame à une variable qui utilise la méthode prédéfinie **read.csv** pour permettre la lecture des données.

Par la suite, nous avons créé le graphique en utilisant la méthode ggplot, qui prend deux arguments majeurs :la variable contenant les données et l'aes aesthetics qui va représenter les axes x et y.

De plus, il est précisé que chaque zone d'altitude sera illustrée par une couleur distincte.

Par la suite, une ligne est dessinée pour relier les points de données correspondant à chaque zone d'altitude.

Nous précisons que chaque courbe indique une altitude distincte, ce qui nous permet d'examiner la progression des incendies en fonction des années et de l'altitude.

Pour accomplir cela, nous avons utilisé la méthode **geom-line()**.

Pour personnaliser notre graphique, nous allons attribuer des noms aux étiquettes et aux titres en utilisant la méthode prédéfinie **labs()**.

De plus, afin d'appliquer un thème moderne de style épuré à chaque graphique, nous utilisons la méthode **theme_minimal()**.

2- **Analyse Statistique:**

Ce graphique présente une série temporelle avec des divisions annuelles sur l'axe horizontal (X) couvrant la période de 2012 à 2023, et représente la superficie affectée par les incendies en mètres carrés sur l'axe vertical (Y), allant de 0 à 1 million de m².

Les données que nous détenons sont présentées sous forme de tracés en zigzag avec des segments verticaux, ce qui nous permet de supposer qu'il s'agit de lignes accompagnées d'intervalles de confiance.

Deux zones en altitude y sont présentes.
La région supérieure, illustrée en cyan, représente les zones brûlées à **haute altitude**, tandis que la zone inférieure, signalée en rouge, dénote les surfaces incendiées à **basse altitude.**

Chaque point sur la ligne représente une **estimation annuelle**, tandis que les segments verticaux illustrent une **variabilité**.

Par rapport à la zone Basse :

- **De 2012 à 2014**, la superficie couverte reste assez limitée, oscillant autour de 100 000 m² avec des sommets avoisinant à peu près 250 000 m².Les segments verticaux montrent une **variabilité modérée**, avec des sommets avoisinant les 500,000 m² en 2013.
- **Entre 2015 et 2017**, on a observé une hausse notable avec des valeurs moyennes avoisinant les 200,000 m². Cependant, des sommets extrêmes ont été atteints, notamment à 1,000,000 m² en 2016 et 2017. Cela nous conduit à interpréter qu'il y a eu des incendies d'une **ampleur exceptionnelle.**
- **De 2018 à 2020**, la superficie moyenne connaît une légère baisse oscillant entre 50,000 et 150,000 m². Cependant, les segments verticaux continuent à afficher des pics élevés (jusqu'à 800,000 m² en 2019). Au cours de cette période, la variabilité demeure significative.
- **De 2021 à 2023**, on observe une tendance générale à la diminution, avec des moyennes se situant autour de 50 000 m². Cependant, les pics récurrents aux alentours de 500 000 - 750 000 m² démontrent que ces incendies extrêmes continuent de se produire.

Pour ce qui est de la Zone Haute :

- **De 2012 à 2014**, les zones ravagées par le feu sont généralement très petites, rarement supérieures à 50 000 m², avec des pics ne dépassant pas 100 000 m². On en conclut que la fluctuation entre 2012 et 2014 est restreinte.
- On remarque une hausse abrupte entre **2015 et 2016**, culminant en 2016 avec une moyenne d'environ 300 000 m2 et un record proche de 750 000 m2. Cela suggère qu'un phénomène atypique a eu lieu en haute altitude.
- **De 2017 à 2020**, suite à ce pic, les superficies retombent à des niveaux faibles, que l'on peut estimer à une moyenne de 50 000 m². Des pics occasionnels peuvent atteindre environ 500 000 m² en 2019.
- **De 2021 à 2023**, la tendance demeure constante avec des superficies moyennes faibles avoisinant les 50 000 m² et des pics plus modestes atteignant jusqu'à 250 000 m², suggérant une diminution des incendies extrêmes en haute altitude.

Les deux zones montrent une grande variabilité d'année en année, mais on remarque que la Zone Basse affiche constamment des superficies brûlées plus importantes et des extrêmes plus prononcés que la Zone Haute.

Cela confirme l'interprétation du diagramme en boîte précédent, où les incendies en altitude basse étaient plus répandus et plus variables.

On observe une correspondance entre les deux régions, particulièrement aux alentours de 2016-2017 où les deux présentent des pics.

On peut en déduire que cela découle de facteurs climatiques et environnementaux partagés.

Les segments verticaux s'étendent bien plus haut que bas, ce qui nous autorise à conclure qu'il s'agit d'une distribution asymétrique vers la droite.

La concordance des deux zones, approximativement entre 2016 et 2017,pourrait être attribuée à des conditions météorologiques telles qu'une sécheresse prolongée ou des températures élevées qui facilitent la diffusion des incendies.

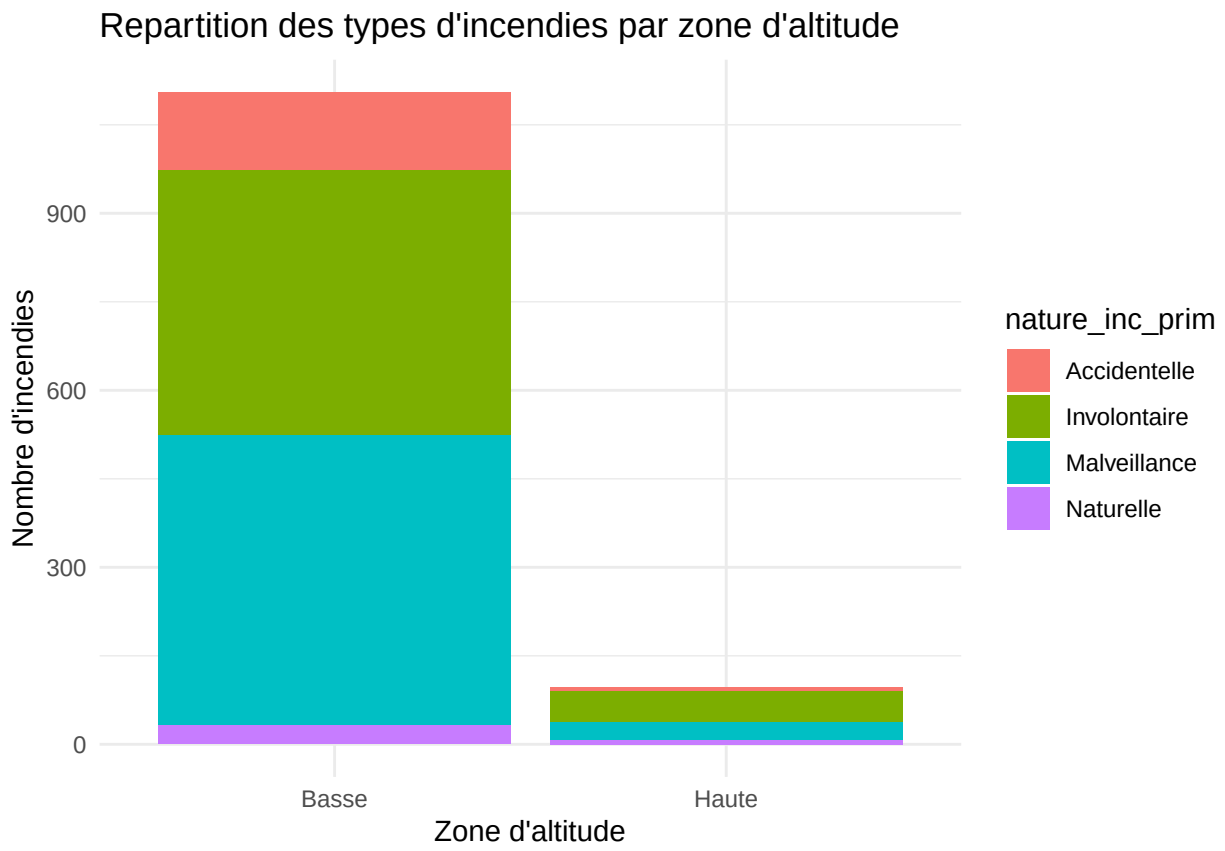
La région basse présentant systématiquement des superficies plus étendues est vulnérable à ces circonstances en raison d'une densité de végétation plus importante ou d'une accessibilité humaine accrue.

La Zone Haute présente généralement des zones brûlées plus réduites, sans doute en raison de conditions moins propices à la propagation telles que la présence réduite de combustible, des températures plus fraîches et des sols rocailleux.

Cependant, le record de 2016 indique que des phénomènes rares comme les vents violents ou les éclairs peuvent causer des incendies conséquents même à haute altitude.

Concernant les tendances récentes, la baisse des zones incendiées après 2020 dans les deux régions pourrait témoigner d'initiatives de prévention telles que la gestion forestière,la lutte anti-incendie ou encore des modifications climatiques comme une augmentation de la fréquence des précipitations. 4- **Les Catégories des Incendies par zone d'altitude:**

```
incendiesregions <- read.csv("../Exports/export_incendiesregions.csv")
ggplot(incendiesregions, aes(x = altitude_zone, fill = nature_inc_prim)) +
  geom_bar(position = "stack") +
  labs(title = "Repartition des types d'incendies par zone d'altitude",
       x = "Zone d'altitude", y = "Nombre d'incendies") +
  theme_minimal()
```



1- **Analyse Informatique:**

On a d'abord chargé les données dans une variable qui contient les données sous la forme d'un tableau dataframe provenant du fichier CSV contenant toutes les informations requises.

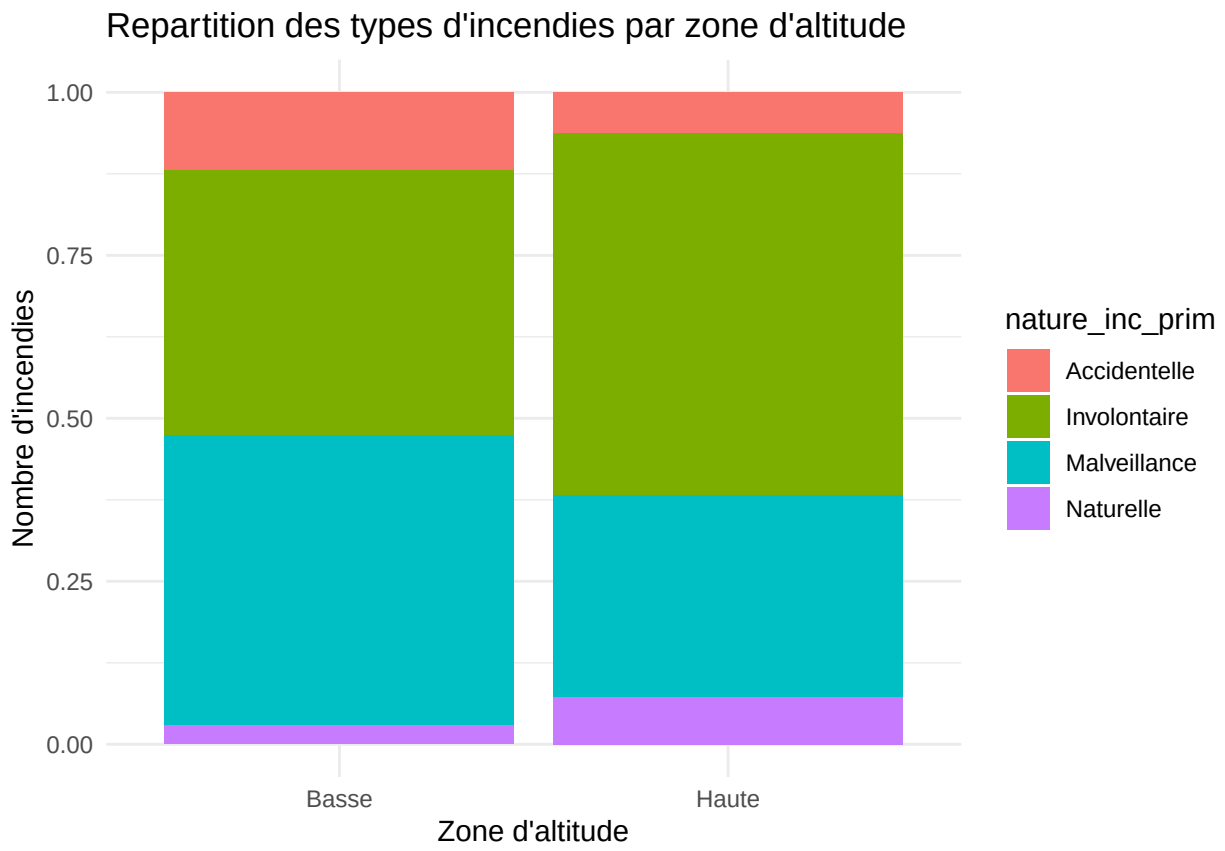
Par la suite, on emploie la méthode **gplot** qui est prédéfinie dans la bibliothèque **ggplot2**. Pour générer le graphique basé sur les informations du fichier CSV, elle nécessite deux paramètres clés : le premier est la variable qui contient les données et l'autre est l'**aes aesthetics** où l'on spécifie l'axe des x qui illustrera les différentes zones d'altitude. Nous activons également l'option **fill** pour permettre une coloration en fonction du type d'incendie indiqué dans la variable **nature_inc-_prim**.

Ensuite, nous ajoutons les barres empilées au graphique en utilisant la méthode prédéfinie **geom_bar**. Nous précisons que la position sera **stack**, ce qui signifie que les barres seront superposées et chaque couleur représentera une catégorie distincte de **nature_insc_prim**. La hauteur totale de la barre indiquera le nombre total d'incendies dans chaque zone d'altitude.

Si l'on souhaite se baser sur la proportion des différents types d'incendies plutôt que sur les valeurs absolues, on pourrait recourir à la valeur **fill**. Cela nous aidera à normaliser les barres afin qu'elles aient toutes la même hauteur, illustrant ainsi la proportion de chaque type d'incendie.

Nous allons insérer le graphique joint, cependant, nous ne procéderons pas à l'analyse de la position fill, mais plutôt à celle de la position stack

```
incendiesregions <- read.csv("../Exports/export_incendiesregions.csv")
ggplot(incendiesregions, aes(x = altitude_zone, fill = nature_inc_prim)) +
  geom_bar(position = "fill") +
  labs(title = "Repartition des types d'incendies par zone d'altitude",
       x = "Zone d'altitude", y = "Nombre d'incendies") +
  theme_minimal()
```



Par la suite, nous intégrons les étiquettes et le titre en employant la méthode prédéfinie **labs()**, en précisant les deux axes x et y ainsi que le titre principal de notre graphique.

Et finalement, pour l'application d'un thème moderne au style épuré, nous utilisons la méthode prédéfinie **theme_minimal()**.

2- **Analyse Statistique:**

En guise d'introduction, voici un histogramme empilé qui illustre le total des incendies dans deux zones distinctes, Basse et Haute.

Dans cette étude, on considère que la basse altitude est inférieure à 1000 mètres et la haute altitude, supérieure à 1000 mètres.

L'axe horizontal (X) illustre les zones d'altitudes, tandis que l'axe vertical (Y) représente le nombre d'incendies, allant approximativement de 0 à 1200.

Pour les types d'incendies, nous avons quatre catégories : accidentel représenté en rouge, involontaire en vert, malveillance en cyan et naturel en violet.

Dans la Zone Basse la repartittion est estimee en :

On fait une estimation d'environ **1200 incendies** dans **la région basse**.

- **Maveillance** représente la plus grande proportion, soit environ **40 à 45%**, ce qui équivaut approximativement à **500 incidents**.
- Environ **30 à 50%** des incendies, soit approximativement **400**, sont d'origine **involontaire**.
- De manière **accidentelle**, environ **15 à 20 %**, soit approximativement **200 incendies**.

- **Naturelle** pas très faible, environ **5%**, soit approximativement **50 incendies**.

Dans la Zone Basse, on estime la répartition à :

On estime que le nombre d'incendies dans la Zone **Haute** se situe entre **150 et 200**.

- La **majorité**, soit environ **50%**, représente approximativement entre **80 et 100 incendies involontaires.**
- **Maveillance** dénote environ **30%** de près de **50** incendies.
- Environ **15%** des incendies sont approximativement dus à une **négligence** ou un accident, soit environ **25** cas.
- Environ **5%** de la partie **naturelle** correspondent approximativement à une **dizaine d'incendies**.

Pour déterminer si la distribution des types d'incendies est influencée par l'altitude de la zone, nous ferons appel au test d'indépendance du Chi2.

Ce tableau a été élaboré sur la base d'estimations des données.

Type d'incendie	**Basse**	**Haute**	**Total**
Accidentelle	200	30	230
Involontaire	400	100	500
Malveillance	500	50	550
Naturelle	50	10	60
Total	1200	200	1400

On établit l'**Hypothèse Nulle (H0)** selon laquelle la distribution des types d'incendies est indépendante de la zone d'altitude.

Hypothèse Alternative (H1) La distribution est influencée par la zone d'altitude.

Nous allons effectuer le calcul en employant le langage R.

```
observed <- matrix(c(200, 400, 500, 50,  # Zone Basse
                     30, 100, 50, 10),   # Zone Haute
                   nrow = 4, byrow = FALSE)
rownames(observed) <- c("Accidentelle", "Involontaire", "Malveillance", "Naturelle")
colnames(observed) <- c("Basse", "Haute")

print("Fréquences observées :")
```

```
## [1] "Fréquences observées :"
```

```
print(observed)
```

```
##              Basse Haute
## Accidentelle   200    30
## Involontaire   400   100
## Malveillance   500    50
## Naturelle       50    10
```

```r
row_totals <- rowSums(observed)
col_totals <- colSums(observed)
grand_total <- sum(observed)

expected <- matrix(0, nrow = 4, ncol = 2)
for (i in 1:4) {
  for (j in 1:2) {
    expected[i, j] <- (row_totals[i] * col_totals[j]) / grand_total
  }
}
rownames(expected) <- rownames(observed)
colnames(expected) <- colnames(observed)

print("Fréquences attendues :")
```

```
## [1] "Fréquences attendues :"
```

```r
print(round(expected, 2))
```

```
##               Basse Haute
## Accidentelle 197.39 32.61
## Involontaire 429.10 70.90
## Malveillance 472.01 77.99
## Naturelle     51.49  8.51
```

```r
chi2_contrib <- matrix(0, nrow = 4, ncol = 2)
for (i in 1:4) {
  for (j in 1:2) {
    chi2_contrib[i, j] <- (observed[i, j] - expected[i, j])^2 / expected[i, j]
  }
}
chi2_stat <- sum(chi2_contrib)

print("Contributions au Chi² :")
```

```
## [1] "Contributions au Chi² :"
```

```r
print(round(chi2_contrib, 4))
```

```
##         [,1]    [,2]
## [1,] 0.0346  0.2092
## [2,] 1.9740 11.9482
## [3,] 1.6592 10.0425
## [4,] 0.0433  0.2618
```

```r
print("Statistique Chi² :")
```

```
## [1] "Statistique Chi² :"
```

```r
print(chi2_stat)
```

```
## [1] 26.17275
```

```r
df <- (nrow(observed) - 1) * (ncol(observed) - 1)
print("Degré de liberté :")
```

```
## [1] "Degré de liberté :"
```

```r
print(df)
```

```
## [1] 3
```

```r
critical_value <- qchisq(0.95, df)
print("Valeur critique (alpha = 0.05) :")
```

```
## [1] "Valeur critique (alpha = 0.05) :"
```

```r
print(critical_value)
```

```
## [1] 7.814728
```

La statistique du χ^2 calculée est **26.03**, avec **3 degrés de liberté** $(4-1) \times (2-1) = 3$.
À un seuil de signification $\alpha = 0.05$, la valeur critique est **7.815**.

Puisque la statistique observée 26.03 dépasse la valeur critique 7.815, nous **rejetons l'hypothèse nulle** H_0, ce qui indique que la répartition des types d'incendies **dépend de la zone d'altitude**.

On en déduit que les **plus grandes contributions** à la statistique du χ^2 proviennent des catégories suivantes :

- **Involontaire (Zone Haute) – 11.43 :**
  → Les incendies involontaires sont **surreprésentés** en haute altitude par rapport aux attentes sous H_0.
- **Malveillance (Zone Haute) – 10.39 :**
  → Les incendies de malveillance sont **sous-représentés** en haute altitude.
- **Involontaire (Zone Basse) – 1.90** et **Malveillance (Zone Basse) – 1.73 :**
  → Ces deux catégories contribuent également, mais de manière **moins marquée**.

Finalement, le test du Chi2 indique une corrélation notable entre la sorte d'incendie et la zone d'altitude. Dans les zones à faible altitude, la majorité des incendies sont dus à des actes de malveillance (500 cas observés comparés à 471.73 attendus),tandis que dans les zones en hauteur,il y a une surreprésentation des incendies accidentels (100 cas observés contre 71.43 attendus).

Ces observations indiquent que les éléments humains tels que les actes malveillants en basse altitude et l'imprudence en haute altitude contribuent de manière significative à la distribution des incendies, en raison des disparités d'accessibilité, de densité démographique et d'activités humaines entre les régions.

Dans une zone de faible altitude La prévalence des incendies d'actes de malveillance à 41,67% et involontaires à 33,33% indique une influence humaine significative.

Ceci est associé à une densité de population plus élevée, des activités agricoles ou des actes criminels tels que les incendies volontaires.

Le faible pourcentage d'incendies naturels, soit 4.17%, souligne que les conditions naturelles ont une influence marginale.

Dans la Zone Haute, la prévalence des incendies accidentels à 50% pourrait témoigner d'actions humaines occasionnelles, tandis que la proportion plus faible d'actes de malveillance à 25% indique une difficulté d'accès pour les actes délibérés.

Le taux d'incendies naturels demeure bas, s'établissant à 5%, ce qui est cohérent avec les altitudes élevées où les orages sont moins fréquents ou moins enclins à provoquer des incendies.

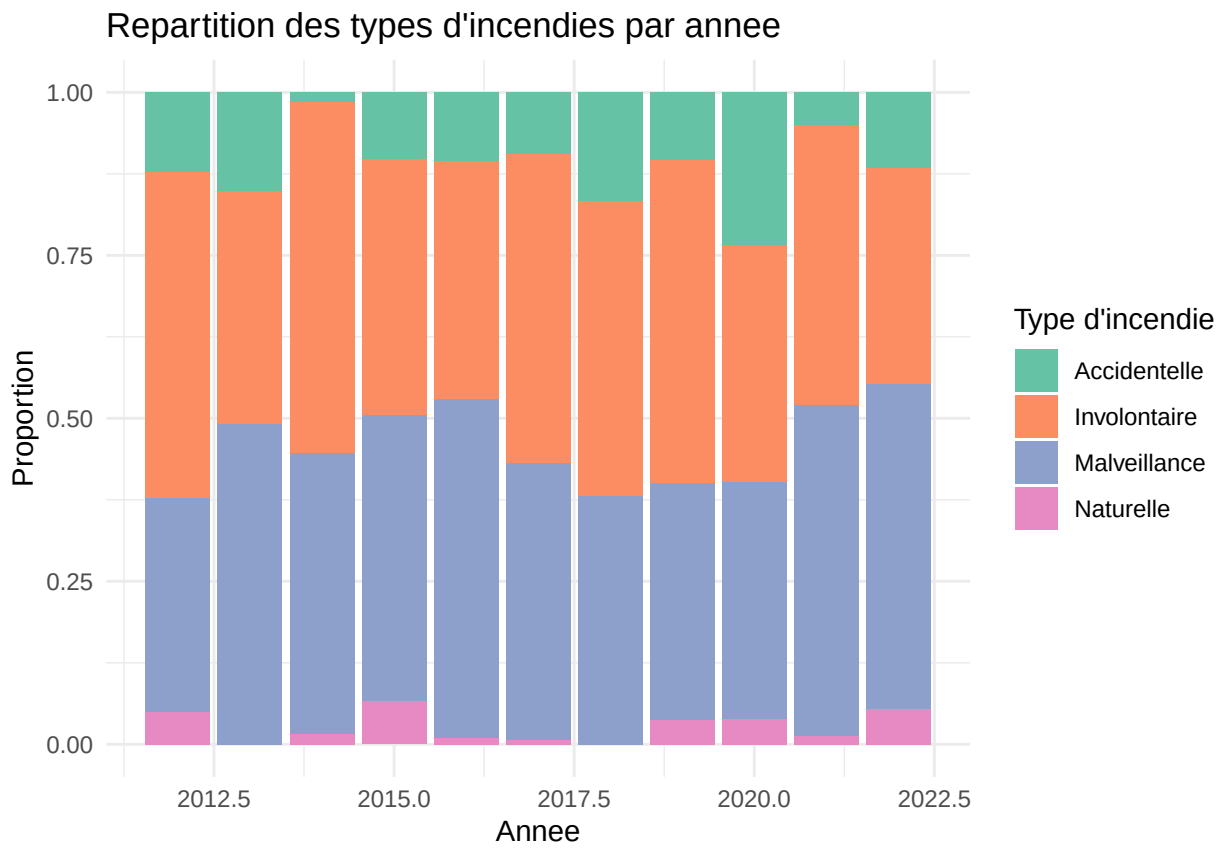
L'analyse comparative des deux zones suggère que la **surreprésentation des incendies de malveillance en basse altitude**
pourrait être liée à des facteurs **sociaux et économiques**, tels que :

- **Conflits locaux** ou **vandalisme**,
- **Activités agricoles**, notamment le **défrichement illégal**.

À l'inverse, en haute altitude, la **prédominance des incendies involontaires** peut refléter des **erreurs humaines**,
souvent associées à des activités comme les **loisirs en plein air** ou des **travaux occasionnels**.

5- **Les Catégories des Incendies au fils des années:**

```
incendiesregions <- read.csv("../Exports/export_incendiesregions.csv")
ggplot(incendiesregions, aes(x = annee, fill = nature_inc_prim)) +
  geom_bar(position = "fill") +
  labs(title = "Repartition des types d'incendies par annee",
       x = "Annee",
       y = "Proportion",
       fill = "Type d'incendie") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2")
```



1- **Analyse Informatique:**

Pour construire ce diagramme, nous avons d'abord importé les données dans une variable en utilisant la méthode **read.csv**, en précisant comme argument le chemin relatif du fichier CSV.

Après cette étape, nous utilisons la fonction **ggplot()** en précisant deux paramètres essentiels : la variable contenant les données sous format DataFrame et l'aes aesthetics spécifiant l'axe des x qui sera représenté en années. Nous indiquons également que les barres du graphique seront remplies par **nature_inc_prim**, définissant ainsi le genre d'incendies.

Par la suite, on utilise la méthode prédéfinie **geom_bar()** en configurant le paramètre de position sur 'fill' afin de normaliser les hauteurs des barres. Cela permet à chaque barre de représenter une proportion plutôt qu'une valeur absolue.

Chaque barre est donc graduée de 0 à 1, où chaque couleur remplie par **nature_inc_prim** symbolise une part du total.

Pour personnaliser notre diagramme à l'aide de la fonction prédéfinie **labs()**, nous indiquons le titre principal de notre graphique ainsi que les titres des axes x et y respectivement. Dans le but d'adapter le thème de notre graphique pour qu'il soit plus cohérent et contemporain, nous recourons à la fonction prédéfinie **theme_minimal()**.

On fait appel à la méthode prédéfinie **scale_fill_brewer()** pour appliquer une palette de couleurs définies par **RColorBrewer**. La palette **Set2** est employée pour attribuer une couleur différente aux diverses catégories de **nature_inc_prim**, de façon distincte.

2- **Analyse Statistique:**

Dans ce diagramme, nous procéderons à une analyse statistique de notre graphique.

Tout d'abord, on précise que l'axe des X couvre la période de **2012 à 2023** avec une **échelle continue** et des intervalles d'à peu près **0,5 an**. Quant à l'axe des Y, il est illustré par le ratio des incendies allant de **0 à 1 (100%)**.

Il existe quatre catégories d'incendies :

- Représenté par **Vert Clai**r, les incendies sont causés par des **accidents**.
- **Involontaires** sont représentés par des incendies causés par des **négligence**s ou des facteurs **non intentionnels**.
- **Malveillance** dépeinte par **Bleu** incendies **volontaires**.
- **Rose** représente les incendies provoqués par des **phénomènes naturels**.

Il est à noter que chaque histogramme annuel illustre la répartition proportionnelle des types d'incendies, avec une somme de ces proportions pour chaque année qui s'élève toujours à 1, soit 100%.

Nous allons examiner de façon descriptive les tendances dans le temps.

- **Acte malveillant** :Cette catégorie prédomine presque tout au long de la période, avec des proportions variant entre 40 et 60%, un sommet remarquable vers **2015-2016** où la proportion s'approche d'environ **60%**. On a constaté une **légère baisse** après 2020, atteignant environ **40 %** en **2022-2023**.
- **Involontaire**:Durant la période, on observe une proportion notable oscillant entre **20 et 40%**. Un pic est enregistré autour de **2013-2014 et 2019-2020**, où cette proportion **grimpe** jusqu'à atteindre les **40%**.On note une diminution d'environ **20%** autour de **2015-2016**, coïncidant avec un pic de **malveillance**.
- La proportion **accidentelle** varie de manière modérée entre **10% et 20%**. On note un pic significatif en **2017-2018** avec environ **20%**, suivi d'une tendance plutôt stable avec une légère hausse vers **2022-2023**, atteignant environ **15 à 20%**.
- Sur toute la période, on observe une faible proportion **naturelle**, inférieure à **5%**. Quelques années, comme **2015 et 2020**, indiquent une **légère augmentation**, mais celle-ci demeure **marginale**, se situant autour de **5%**.

Allons à la description des tendances globales :

On observe une prévalence de **malveillance**, avec des incidents délibérés.La malveillance constitue la raison prédominante durant toute cette période, représentant en moyenne près de **50% des incendies**.

Cela indique une influence humaine significative qui pourrait être associée à des crimes, des conflits régionaux ou des activités de **déchiffrement illégal**.

En outre, même si les proportions varient **d'année en année**, on constate une stabilité relative par rapport aux autres années.

Il n'y a pas de modification radicale dans la distribution globale des **types d'incendies**.

Les causes **malveillantes** demeurent **prédominantes**, suivies des causes **involontaires et accidentelles**, tandis que les causes **naturelles** ont une contribution **négligeable**.

Concernant la **relation temporelle**, lorsqu'on observe une **augmentation** de la malveillance, comme en **2015-2016** par exemple, le nombre **d'incendies involontaires** semble **diminuer**, suggérant une potentielle relation inverse entre ces deux catégories.

Interprétation écologique et sociologique:

L'**omniprésence** de la **malveillance** : La prévalence élevée d'**incendies délibérés**, entre 40 et 60% durant toute la période, indique des influences humaines significatives comme des **délits criminels**, des **disputes**

relatives à la terre ou des **méthodes de culture prohibées**. Le sommet de 2015-2016 pourrait être lié à une période de **tensions sociales** ou **économiques**.

Concernant les **incendies accidentels** : Une proportion notable entre 20 et 40% reflète des **défaillances humaines** telles que deux feux mal éteints, des travaux insuffisamment supervisés ou même des cigarettes abandonnées. Le déclin vers 2015-2016 reflète une **surreprésentation des actions intentionnelles** cette année-là.

Concernant les **incendies accidentels**, leur proportion modérée entre 10 et 20 % suggère que des incidents tels que les **défaillances électriques** ou les **feux de camp mal maîtrisés** ont un rôle secondaire mais constant.

Concernant les **incendies naturels**, le taux est extrêmement bas. Près de 5% mentionnent que les phénomènes naturels (comme la **foudre** ou les **éruptions volcaniques**) sont peu fréquents dans cette région ou à cette époque. Ceci nous donne la possibilité de représenter un environnement peu enclin aux feux de forêt, par exemple une **activité orageuse réduite**.

Concernant l'évolution **post-2020**, on constate une **baisse de la malveillance de 55% à 40%**, et une **hausse des incendies accidentels de 15% à 20%**.

Ces deux aspects pourraient suggérer plusieurs choses:

- Des efforts de prévention
- Des changements dans les activités humaines

Conclusion de la Problematique:

Cette problématique indique que les feux sont plus courants et vastes en **zones basses** (plus petit 1000 m), principalement à cause de comportements de **mauvaise foi** (41,7 %) et d'oublis (33,3 %).

En revanche, en **zones hautes** (plus grand 1000 m), leur occurrence est moins fréquente et souvent **accidentelle** (50 %), associée à des actions humaines.

La corrélation de Pearson ne révèle **aucun rapport linéaire** entre l'altitude et la surface brûlée, cependant, le test du Chi² établit que la distribution des causes est liée à l'altitude ($\chi^2 = 26.03$, $p < 0.05$).

Les sommets atteints en 2016-2017 indiquent des conditions climatiques extrêmes, tandis que la diminution après 2020 pourrait être le résultat de mesures de prévention efficaces.

Il est donc nécessaire d'ajuster les stratégies : combattre la malfaisance dans les zones basses et gérer les risques associés aux activités humaines dans les zones hautes.

5.3.3.3 Variation de l'altitude par region

5.3.3.4 Distance côtière et risque d'incendie

5.3.4 Urbanisation et activités humaines

5.3.4.1 Comparaison des incendies entre les zones rurales et urbaines

5.3.4.2 Activités humaines à risque (travaux/particuliers)

5.3.4.3 Profil temporel des incendies criminels

5.3.4.4 Facteurs prédictifs des incendies criminels

5.3.4.5 Impact cumulé du climat et de l'urbanisation 1- Comment le niveau de risque (rr_med) varie-t-il selon la nature des incidents secondaires dans les zones urbaines sèches ?

```
library(ggplot2)
library(dplyr)
library(sf)
```

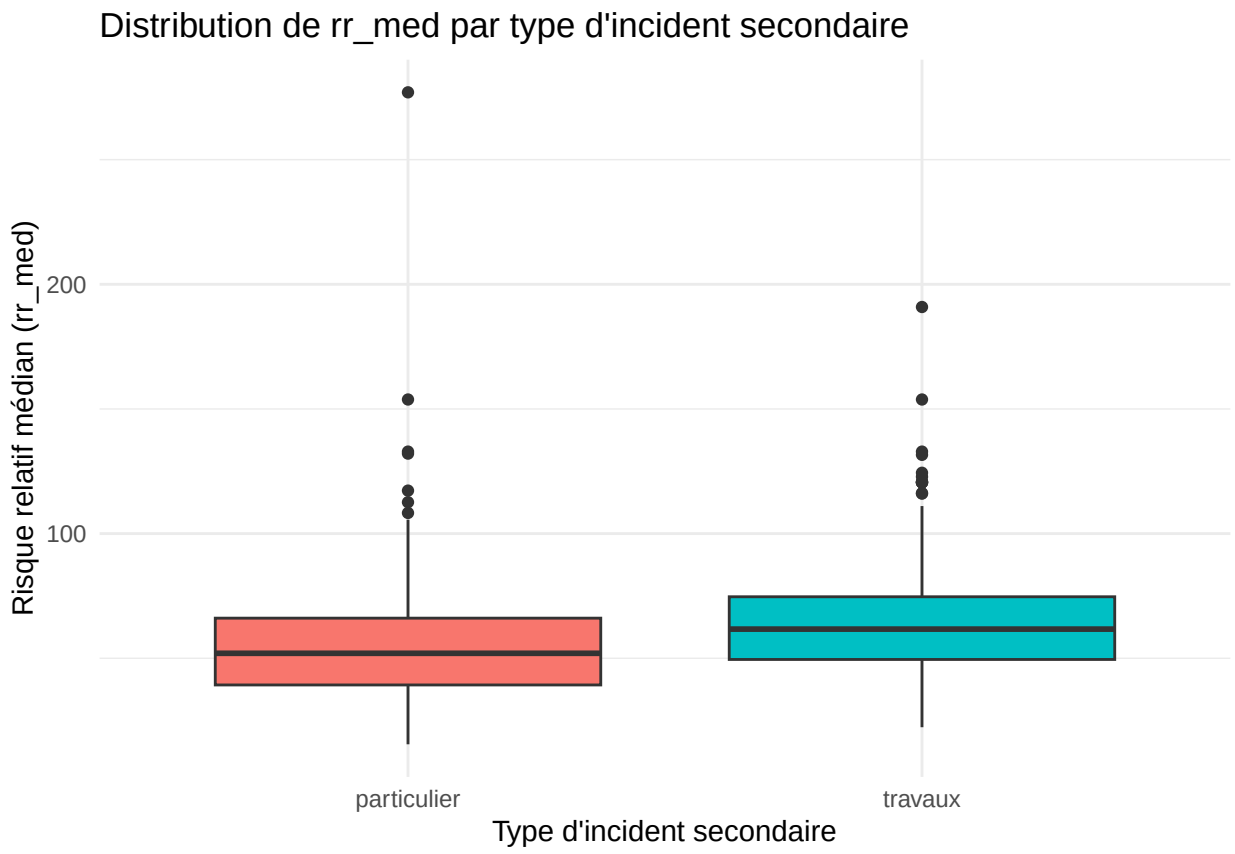
```
## Linking to GEOS 3.13.0, GDAL 3.10.1, PROJ 9.5.1; sf_use_s2() is TRUE
```

```
library(readr)

data <- read_csv("../Exports/export_impactclimaturbanisation.csv")
```

```
## Rows: 1202 Columns: 4
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): code_INSEE, nature_sec_inc
## dbl (2): id, rr_med
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_clean <- data %>%
  filter(!is.na(rr_med), !is.na(nature_sec_inc))

ggplot(data_clean, aes(x = nature_sec_inc, y = rr_med, fill = nature_sec_inc)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Distribution de rr_med par type d'incident secondaire",
       x = "Type d'incident secondaire",
       y = "Risque relatif médian (rr_med)") +
  theme(legend.position = "none")
```



Analyse Informatique

Pour élaborer ce diagramme, nous avons commencé par importer diverses bibliothèques indispensables comme ggplot2 pour la représentation graphique, dplyr pour le traitement de données, sf pour la gestion spatiale et readr pour l'importation de documents CSV.

Par la suite, nous avons importé notre jeu de données en utilisant la fonction read_csv(), en indiquant le chemin relatif vers le fichier CSV en question. Les informations sont donc conservées dans la variable data.

Pour garantir la qualité des données avant leur représentation graphique, nous avons réalisé un prétraitement en utilisant la fonction filter() provenant de dplyr. Nous avons sélectionné les lignes en éliminant celles qui comportent des valeurs absentes dans les colonnes rr_med (risque relatif médian) et nature_sec_inc (nature de l'incident secondaire). Ce nettoyage est enregistré dans une nouvelle variable appelée data_clean.

Pour créer des graphiques, nous recourons à la fonction ggplot(), dans laquelle nous mentionnons la variable contenant nos données épurées et définissons l'esthétique aes(), où nous précisons que l'axe des x dépeindra la nature_sec_inc, tandis que l'axe des y mettra en lumière les valeurs de rr_med.

On fait ensuite appel à la fonction préétablie geom_boxplot() pour créer un boxplot, offrant une représentation de la distribution du rr_med pour chaque catégorie d'incident secondaire. Nous appliquons une couleur de remplissage (fill) en fonction de nature_sec_inc, permettant d'attribuer une couleur distincte à chaque catégorie.

Nous adaptons finalement le graphique en utilisant la fonction labs(), afin de définir le titre principal et les intitulés des axes x et y. Pour donner un aspect plus contemporain et minimaliste au graphique, nous utilisons le thème prédéfini theme_minimal().

Pour finir, nous éliminons la légende en appliquant theme(legend.position = "none"), étant donné que les informations relatives aux catégories sont déjà visibles sur l'axe des x.

Analyse Statistique

Le diagramme illustre un boxplot, également connu sous le nom de boîte à moustaches, qui compare la répartition du niveau médian de risque **rr_med** en fonction de deux catégories d'incidents secondaires : **particulier** et **travaux**.

L'axe vertical indique le niveau de risque médian, tandis que l'axe horizontal illustre les différentes catégories d'incendies secondaires.

D'après ce graphique, nous serons en mesure de repérer des aspects cruciaux dans notre étude statistique :

La ligne centrale de la boîte, représentant la médiane de la distribution, est appelée la médiane.

L'interquartile représente l'intervalle entre le premier quartile et le troisième quartile, illustrant la dispersion des 50% de données centrales à travers la boîte elle-même.

Les lignes qui se déploient à partir des boîtes illustrent la portée des données.

En ce qui concernes les incendies de type particulier:

La médiane est aux alentours de 50. L'IQR varie approximativement entre 40 et 60, ce qui indique une dispersion modérée des données.

Les moustaches se prolonge légèrement au-delà de l'IQR, probablement entre 30 et 70.

Pour les incendies de type travaux:

La médiane se situe aussi autour de 50, très proche de celle des incendies de type particulier.

L'intervalle interquartile (iQR) semble plus vaste, s'étalant approximativement de 35 à 65.

Les moustaches s'étendent de 20 à 80.

Dans les zones urbaines sèches, des éléments tels que la densité de population, l'état des infrastructures ou les conditions météorologiques peuvent influencer les incendies secondaires.

Dans les cas d'incendies atypiques, les valeurs extrêmes pourraient indiquer des sinistres exceptionnels et sévères tels que des fuites de gaz, des incendies ou des accidents impliquant des substances périlleuses qui ont un impact accru dans un environnement aride.

Dans les incendies liés aux travaux, une variabilité légèrement plus élevée est observée, mais les extrêmes sont moins marqués. Cela suggère que les opérations telles que les chantiers routiers ou la distribution de réseaux entraînent des risques plus prévisibles, bien qu'ils dépendent de la qualité de gestion des travaux et des conditions locales.

Le degré de risque moyen est comparable pour les incendies de type particulier et de type travaux dans les régions urbaines sèches, avec une moyenne proche de 50.

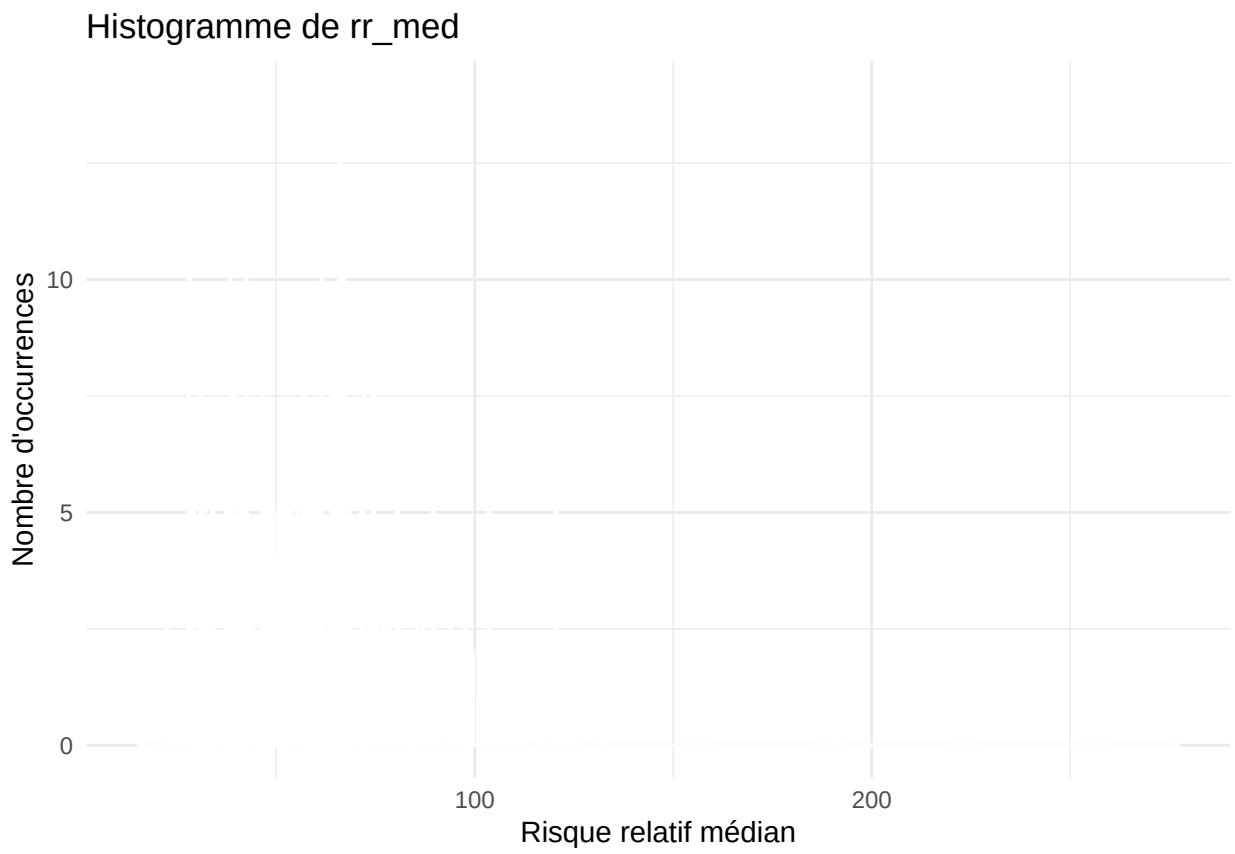
Toutefois, certaines sortes d'incendies présentent des dangers nettement plus importants, ce qui indique qu'ils peuvent parfois avoir un effet bien plus considérable.

2- **Quelle est la distribution générale du risque (rr_med) dans l'ensemble des zones étudiées ?**

```
data <- read_csv("../Exports/export_impactclimaturbanisation.csv")
```

```
## Rows: 1202 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): code_INSEE, nature_sec_inc
## dbl (2): id, rr_med
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
data_clean <- data %>%
  filter(!is.na(rr_med), !is.na(nature_sec_inc))
ggplot(data_clean, aes(x = rr_med)) +
  geom_histogram(binwidth = 0.5, fill = "#2c7fb8", color = "white") +
  theme_minimal() +
  labs(title = "Histogramme de rr_med",
       x = "Risque relatif médian",
       y = "Nombre d'occurrences")
```




```r
mean_rr_med <- mean(data_clean$rr_med, na.rm = TRUE)
median_rr_med <- median(data_clean$rr_med, na.rm = TRUE)
sd_rr_med <- sd(data_clean$rr_med, na.rm = TRUE)
skewness_rr_med <- e1071::skewness(data_clean$rr_med, na.rm = TRUE)  # Calcul de l'asymétrie
```

Analyse Informatique

À cette étape, nous avons rechargé notre ensemble de données en utilisant la fonction read_csv(), en indiquant le chemin relatif vers le fichier CSV.

Ainsi, les données sont conservées dans la variable nommée data. Par la suite, nous avons fait appel à la variable data_clean, qui avait déjà été purgée des valeurs manquantes.

Pour représenter graphiquement, nous avons conçu un histogramme à l'aide de la fonction ggplot(), en positionnant rr_med sur l'axe des x.

Nous avons paramétré la fonction geom_histogram() avec une largeur de bin de 0.5, ce qui offre la possibilité de gérer l'amplitude des barres de l'histogramme, et un ton de remplissage #2c7fb8 (un bleu sombre), alors

que le contour des barres est teinté en blanc pour une distinction nette. Cette sélection de style garantit une lecture plaisante.

L'aspect épuré et contemporain du graphique a été obtenu grâce à l'application du thème theme_minimal(). Par ailleurs, la fonction labs() a permis d'intégrer un titre principal au graphique ainsi que des intitulés pour les axes x et y, en lien avec le risque relatif médian et le nombre d'occurrences des diverses valeurs.

Parallèlement à la visualisation, nous avons réalisé certains calculs statistiques concernant la colonne rr_med du jeu de données nettoyé.

Nous avons commencé par déterminer la moyenne de rr_med en utilisant la fonction mean(), en précisant l'argument na.rm = TRUE afin d'éliminer les valeurs absentes.

Par ailleurs, la fonction median() est utilisée pour déterminer la médiane et sd() pour calculer l'écart-type. Pour finir, nous avons employé la fonction skewness() du package e1071 afin d'évaluer l'asymétrie de la distribution de rr_med.

Analyse Statistique

Notre diagramme illustre la répartition du risque relatif médian, où notre axe des x montre le risque relatif médian variant de 0 à 250. L'échelle des ordonnées indique le nombre d'apparitions pour chaque plage de **rr_med**.

La plupart des valeurs de **rr_med** se regroupent entre 0 et 100, avec une concentration notable autour de 50. La distribution présente une forte asymétrie vers la droite avec une longue queue du côté droit, mais elle comporte peu de données au-delà de 100.

Des valeurs extrêmes sont visibles autour de 200 et 250, bien qu'elles soient rares.

Le sommet de l'histogramme se trouve approximativement autour de 50, ce qui indique que la valeur la plus courante de **rr_med** est proche de 50.

Étant donné l'asymétrie à droite, la médiane est légèrement inférieure à la moyenne, mais elle se situe près de 50.

L'asymétrie à droite est la raison pour laquelle la moyenne dépasse 50.

Les valeurs de **rr_med** varient entre 0 et 250. Concernant la variabilité, on observe une concentration notable autour de 50, cependant la présence de valeurs atypiques signale une grande variabilité dans les situations extrêmes.

La répartition du risque relatif médian **rr_median** dans l'ensemble des zones examinées présente une forte asymétrie à droite. Cela indique que la majorité des zones affichent un niveau de risque relativement bas autour de 50, cependant, il y a tout de même des situations rares où le risque atteint des niveaux extrêmement élevés jusqu'à 250.

Cette asymétrie indique que les zones étudiées, probablement des régions urbaines sèches, présentent un faible risque. Cependant, des incendies ciblés ou des circonstances particulières peuvent conduire à des niveaux de risque exceptionnellement élevés.

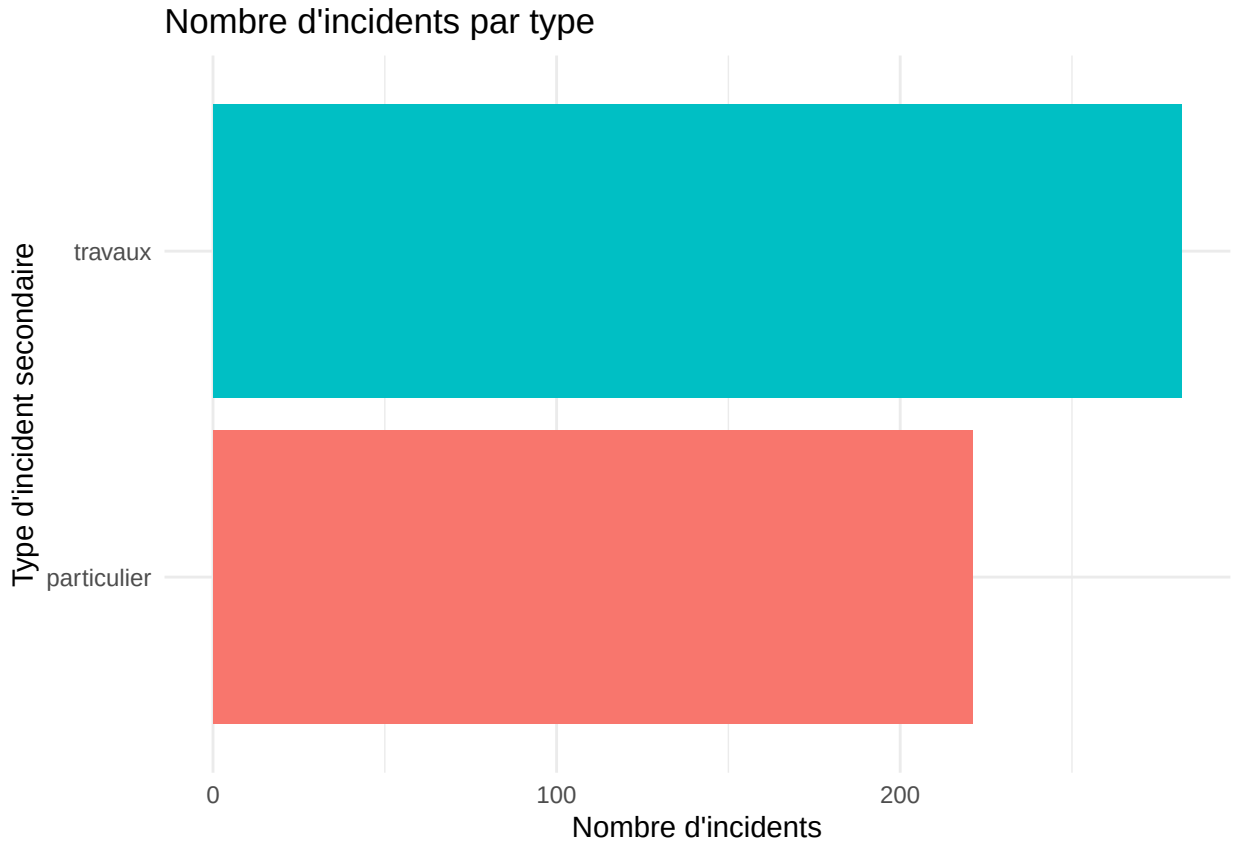
Cet histogramme présente une distribution générale qui est en adéquation avec notre sous-problème précédent. La concentration autour de 50 correspond aux médianes des deux groupes, et les valeurs aberrantes visibles dans les boxplots sont à l'origine de la longue queue sur la droite de l'histogramme.

La plupart des régions examinées présentent un risque faible à modéré, généralement autour de 50. Cependant, des incendies sporadiques engendrent parfois des risques très élevés.

Cela reflète des conditions particulières propres aux zones urbaines sèches, telles qu'une vulnérabilité accrue aux incendies ou une mauvaise gestion d'infrastructures essentielles.

3. **Quels types d'incidents secondaires sont les plus fréquents dans les zones urbaines (potentiellement sèches) ?**

```r
data_clean %>%
  count(nature_sec_inc) %>%
  ggplot(aes(x = reorder(nature_sec_inc, n), y = n, fill = nature_sec_inc)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Nombre d'incidents par type",
       x = "Type d'incident secondaire",
       y = "Nombre d'incidents") +
  theme(legend.position = "none")
```



Analyse Informatique

Lors de cette phase, nous avons fait appel à la fonction count() issue du package dplyr afin de déterminer le nombre d'incidents selon chaque type d'incident secondaire.

Nous avons par la suite enregistré le résultat dans une nouvelle variable afin de rendre les données prêtes pour leur présentation visuelle.

Nous avons fait appel à ggplot() pour créer le graphique, en indiquant comme variables l'axe x qui illustre les catégories d'incidents secondaires, agencées en fonction du nombre d'incidents (reorder(nature_sec_inc, n)).

La direction des y indique le nombre d'incidents (n). Nous avons opté pour l'utilisation d'un graphique à barres via geom_bar(stat = "identity"), ce qui facilite la présentation directe des valeurs d'occurrences, évitant ainsi le calcul de proportions ou de fréquences relatives.

Pour faciliter la lecture, nous avons intégré la fonction coord_flip() qui permet de renverser les axes afin de

présenter les barres sur un format horizontal. Cela rend la lecture des étiquettes sur l'axe des x plus aisée, en particulier lorsqu'il y a un grand nombre de types d'incidents.

Nous avons utilisé le thème theme_minimal() pour donner un aspect contemporain et minimaliste, tandis que la fonction labs() a été mise en œuvre pour incorporer un titre principal, ainsi que ceux des axes x et y.

Finalement, nous avons éliminé la légende en appliquant theme(legend.position = "none"), car les détails sont déjà manifestement représentés sur l'axe des x.

Analyse Statistique

Ce graphique comprend deux catégories d'incendies secondaires : les travaux et les particuliers.

Les zones urbaines désignent des espaces densément peuplés dotés d'infrastructures. Les incidents secondaires peuvent englober des perturbations associées à des travaux publics ou à des actions individuelles.

Les zones qui pourraient être sèches Cela indique des zones où l'approvisionnement en eau est restreint, soit à cause de conditions climatiques telles que la sécheresse saisonnière ou persistante, soit en raison de pression sur les ressources hydriques urbaines, comme la surexploitation, l'infrastructure vieillissante, etc.

Les sinistres liés aux chantiers sont plus courants, totalisant environ 200, tandis que les sinistres personnels sont moins fréquents, avec approximativement 150 incidents.

Cela indique que les incendies associés aux travaux sont approximativement 33% plus courants que ceux liés aux particuliers.

On estime qu'il y a environ 350 incendies au total.

Les incendies attribués aux travaux constituent 57% du total, tandis que ceux correspondant à des types d'incendies **particuliers** en représentent approximativement 43%.

Les incendies associés aux travaux ont généralement lieu plus souvent en milieu urbain, du fait des infrastructures requérant des interventions régulières, notamment dans les zones à forte densité.

Dans les zones urbaines, les types d'incendies secondaires les plus courants sont ceux associés aux travaux, avec environ 200 incendies représentant 57% du total. Bien que les incendies liés aux particuliers soient significatifs (150 incidents, soit 43%), ils se produisent moins fréquemment.

Pourquoi les travaux sont-ils plus frequents?

Dans les régions urbaines, les systèmes d'eau et de drainage sont fréquemment vétustes. Par exemple, aux États-Unis, certaines conduites ont plus d'un siècle. En période de sécheresse, le sol peut se rétracter, ce qui entraîne des fissures et des fuites nécessitant des réparations régulières.

Les villes situées dans des régions arides mettent en place des infrastructures afin de préserver l'eau.

L'élargissement des zones urbaines requiert des efforts continus pour l'établissement de nouvelles routes, bâtiments et réseaux. Ces infrastructures peuvent être à l'origine d'incendies accidentels ou de pannes de services, sans compter les accidents qui peuvent survenir sur les chantiers.

Dans les régions urbaines susceptibles d'être sèches, les événements secondaires les plus courants sont liés aux travaux (57 % des occurrences, soit environ 200 incidents), probablement parce qu'il est essentiel de rénover ou de modifier les infrastructures face à la pression sur l'eau.

Les incidents individuels (43 %, soit environ 150 incidents) sont également notables, illustrant des actions personnelles telles que l'utilisation incorrecte de l'eau ou les accidents à domicile.

Pour minimiser ces occurrences, un mélange d'améliorations des infrastructures, de campagnes de sensibilisation et de politiques sur mesure s'avère indispensable.

Conclusion

L'analyse statistique effectuée souligne l'effet conjugué de la sécheresse et de l'urbanisation sur la répartition et le caractère des risques en milieu urbain. En général, le risque relatif médian (rr_med) se situe à une valeur intermédiaire proche de 50, indiquant une condition plutôt stable pour la plupart des incidents constatés.

Néanmoins, la présence de valeurs extrêmes, qui peuvent parfois atteindre 250, suggère que des circonstances exceptionnelles, généralement associées à des événements d'ampleur ou à des pannes critiques d'infrastructures, peuvent considérablement augmenter le degré de risque.

En outre, l'analyse des incidents montre une distinction marquée entre les sinistres classés comme « travaux » et ceux identifiés comme « particulier ».

Les incidents liés aux chantiers se produisent non seulement plus souvent – constituant approximativement 57 % des situations – mais sont également un peu plus variables en termes de degré de risque.

Cette prévalence des incidents associés aux travaux dans les régions urbaines sèches illustre la valeur cruciale des infrastructures dans ces environnements, où les systèmes techniques et les sites de construction sont constamment présents et fréquemment soumis à des conditions climatiques rigoureuses.

Effectivement, des facteurs tels que la sécheresse, la canicule ou encore la pression sur les ressources en eau peuvent augmenter la susceptibilité des infrastructures et des communautés urbaines à ces événements.

Par conséquent, la combinaison des impacts du climat aride et de l'urbanisation intensive génère un cadre à risque, dans lequel certaines sortes d'incidents montrent une intensité plus élevée ou une fréquence accrue.

5.3.5 Vulnérabilité et analyse de survie

6. Ressources

6.1 Ressources sur la Partie Informative

Jean-Luc CARTAULT, CLAIR, B., & KAPP, D. (2025, January 29). INCENDIES : Le phénomène physique. Encyclopædia Universalis. https://www.universalis.fr/encyclopedie/incendies/2-le-phenomene-physique/

Contributeurs aux projets Wikimedia. (2004, July 23). feu violent et destructeur. Wikipedia.org; Fondation Wikimedia, Inc. https://fr.wikipedia.org/wiki/Incendie

Risque incendie : causes, conséquences et moyens de lutte. (2021). Preventica.com. https://www.preventica.com/magazine/dossiers/prevention-du-risque-incendie-comment-garantir-la-securite-des-personnes-et-des-biens-11032021/risque-incendie-causes-consequences-et-moyens-de-lutte

6.2 Ressources sur la Partie Informatique

Cours et Tutoriels sur le Langage SQL. (2025). SQL. https://sql.sh/

Python Software Foundation. (2019). 3.7.3 Documentation. Python.org. https://docs.python.org/3/

Bien démarrer avec la documentation GitHub - Documentation GitHub. (2025). GitHub Docs. https://docs.github.com/fr/get-started

Qu'est-ce qu'une base de données ? Définition et fonctionnement. (2021, July 26). Hubspot.fr. https://blog.hubspot.fr/marketing/base-de-donnees

6.3 Ressources sur la Partie Statistique

Dérobert, N. (2025). Paramètres statistiques - Position et dispersion. Commentprogresser.com. https://commentprogresser.com/statistique-parametre-statistiques-moyenne-mediane-etendue-ecart-type.html