

Membres du groupe

**Tuteur enseignant :** Michaël Corbeau

| **M1** | **M2** |
| --- | --- |
| DOUX Luka  KENGURUKA Alain  LOUNES Farouk Amrane  MERVIN Moana  PY Célia | DIXIMIER François  LABREUR Thomas  VINCENT Lucas |

Gestion des versions

| **Version** | **Description** | **Date** | **Auteur** |
| --- | --- | --- | --- |
| *Version 1.1* | Page de garde | 06/01/2025 | Thomas LABREUR |
| *Version 1.3* | Eléments de rédaction  et Remerciements | 07/01/2025  -  17/01/2025 | Entièreté du groupe |
| *Version 1.4* | Rédaction des paragraphes | 09/01/2025  -  17/01/2025 | Entièreté du groupe |
| *Version 1.5* | Insertion des figures | 15/01/2025 | Entièreté du groupe |
| *Version 2.0* | Finalisation du rapport | 17/01/2025 | Lucas VINCENT, François DIXIMIER, Thomas LABREUR |
| *Version 2.1* | Relecture du rapport | 17/01/2025 | Lucas VINCENT, François DIXIMIER, Thomas LABREUR |

Préambule

Le présent document expose les activités réalisées par le groupe 3 tout au long du projet, mettant en lumière les divers défis liés à la gestion de projet. Les détails du travail accompli sont également consignés de manière plus directive dans les notebooks.

Si votre intérêt se limite à la compréhension et à l'exploitation du travail effectué, il peut être préférable de privilégier les notebooks qui occultent la gestion de projet tout en illustrant plus précisément les méthodes.

Tous les codes développés sont consultables dans notre [dépôt github](https://github.com/Thomas-labreur/PIP2025_GR3.git).

Remerciements

Le projet a pu être réalisé grâce à la collaboration et à l'engagement de différentes personnes, que nous tenons à remercier chaleureusement.

En premier lieu, nous exprimons notre reconnaissance envers l'Université Paul Sabatier qui nous a permis de réaliser ce projet inter-promo au cours de ces deux semaines.

Nous remercions nos tuteurs pédagogiques, Monsieur Michaël CORBEAU ainsi qu’Emmanuelle CLAYES. Leurs conseils et leurs partages ont été une aide précieuse pour mener à bien ce projet.

Nous souhaitons également saluer tous les membres de notre groupe pour leurs contributions. La collaboration en équipe a été une expérience enrichissante tant sur le plan personnel que professionnel, et nous sommes reconnaissants pour l'investissement, le soutien et les apports de chacun.

Enfin, nos remerciements s'adressent à data.gouv.fr de nous avoir offert l'opportunité de réaliser ce projet. La mise à disposition de données qualitatives en open data est une initiative que nous apprécions, car elle nous permet un support fiable pour travailler convenablement.

Table des matières

[**Introduction 6**](#_heading=h.tyjcwt)

[**I. Gestion de projet 8**](#_heading=h.3dy6vkm)

[1. Environnement de travail 8](#_heading=h.exosnv8ljvec)

[2. Gestion du groupe 8](#_heading=h.30y4cnkaoplr)

[3. Problèmes rencontrés 9](#_heading=h.c2fa1cq7eneb)

[4. Gantt du projet 10](#_heading=h.mvgks2q8a1cm)

[Figure 1 : Diagramme de GANTT prévisionnel 10](#_heading=h.an38c8fyttct)

[Figure 2 : Diagramme de GANTT réel 11](#_heading=h.hv8aev7hx4ek)

[**II. Description des données 12**](#_heading=h.1t3h5sf)

[1. Contexte 12](#_heading=h.6qbg3cyq27ns)

[2. Exploration des données 12](#_heading=h.57v2bmoq1bwv)

[Figure 3 : Exemple de contenu d’un fichier csv 13](#_heading=h.qpnwpaev6mc5)

[**III. Méthodologie 14**](#_heading=h.4d34og8)

[1. Méthodologie de création des tableaux de bord 14](#_heading=h.w75ijwpvawhm)

[2. Présentation des premiers tableaux de bord: Thématiques 14](#_heading=h.x6sdikjl8xp)

[Figure 3 : Maquette du premier dashboard 15](#_heading=h.rmwuzvvryaas)

[2.1 Analyse des films: 17](#_heading=h.yuddtzoht4k6)

[2.1.1. Introduction: 17](#_heading=h.621via5hd64f)

[2.1.2 Nettoyage des Données: 18](#_heading=h.hpwii58ealbh)

[2.1.3. Visualisations et Analyses: 18](#_heading=h.z2k86zalue9)

[2.2.Visualisations: 20](#_heading=h.y46ycgckimhq)

[3. Présentation du second tableau de bord: Temps de parole des femmes 23](#_heading=h.rbd8coimslj6)

[4. Algorithme de clustering concernant les chaînes de télévision française 28](#_heading=h.6vf3onbi7icq)

[5. Présentation du troisième tableau de bord: Politique 28](#_heading=h.3hh3d6b1hmie)

[6. Présentation du quatrième tableau de bord: L’âge 30](#_heading=h.7ppaj9yfftub)

[**IV. Résultats 32**](#_heading=h.2s8eyo1)

[1. Les thèmes abordés lors des JT 32](#_heading=h.1peu02507u9f)

[2. Le temps de parole des femmes à la télévision et à la radio 32](#_heading=h.7r3av525j02o)

[3 Les médias et l’opinion politique 38](#_heading=h.4gg2sntlp9wj)

[**V. Conclusion 39**](#_heading=h.pfb5km7kramb)

[1. Conclusion 39](#_heading=h.26in1rg)

[2. Bilan 40](#_heading=h.lnxbz9)

# Introduction

Chaque année, pendant les deux premières semaines de janvier, les étudiants en master 1 et master 2 de la formation SID s'engagent dans le projet inter promo, une initiative qui favorise la collaboration interdisciplinaire et la mise en pratique des compétences acquises.

Pour l'édition de 2025, ce projet prend forme de la participation des deux promotions de SID à des défis portant sur des données publiques fournies par le gouvernement. La thématique retenue pour notre groupe est “Les Françaises et Français face à l’information”

Nous avIons décidé de porter une étude selon trois axes:

* L’évolution des thèmes abordés par les principales chaînes télévisées en France
* L’évolution du temps de parole des femmes dans les médias
* Analyse de proximité de ces chaînes

Nous voulions initialement traiter les deux premiers thèmes via des Dashboard, et le troisième via des algorithmes de clustering. Comme nous l’expliciterons plus tard, l'exploitabilité (ou non) de certaines données et l’ouverture à des données extérieures nous a amené à faire évoluer cette vision en cours de route. En effet, en ce qui concerne le temps de parole des femmes, nous avons voulu également inclure des données sur la représentation des femmes dans les films diffusés sur ces chaînes. Nous avons également abandonné l’idée de clustering par manque de données (5 chaînes de TV seulement), et préféré ouvrir deux nouveaux axes, également traités par dashboards avec une parti machine learning (ML).

Les axes finaux sont donc:

* L’évolution des thèmes abordés par les principales chaînes télévisées en France
* La représentation des femmes dans les médias et les films diffusés à la tv
* La consommation de l’audiovisuel selon l’âge
* L'impact de la télévision sur les opinions politiques (avec ML).

En ce qui concerne la Data Visualisation (les différents dashboard produits), nous avons décidé de ne pas utiliser de solutions non libres telles que PowerBI, Tableau ou bien Apex. Nous nous sommes finalement tournés vers l’utilisation du langage python, muni de la bibliothèque “Gradio”, qui nous permet de produire des visuels interactifs. Le module “Dash” a également été utilisé pour certains visuels.

Chaque groupe a travaillé sur son sujet, cependant comme tous les groupes développent une solution sous forme de dashboard, une forte entraide a pu être mise en place .

Ce rapport détaille notre démarche méthodologique, les approches employées, les calculs de risque effectués, les méthodes d'association, ainsi que l'identification des bâtiments présentant les risques les plus élevés. Il sert de compte-rendu exhaustif du projet, mettant en lumière la gestion du groupe, le travail accompli en adéquation avec les objectifs fixés, et propose une réflexion approfondie sur le déroulement du projet au cours des deux semaines du projet inter-promo 2023.

# Gestion de projet

Dans la partie Gestion de projet de ce rapport, nous examinerons l'environnement de travail, la collaboration au sein du groupe, les problèmes rencontrés, les attentes initiales par rapport à la réalité du projet, ainsi que le diagramme de Gantt illustrant la planification et la réalisation du projet.

## Environnement de travail

Pendant la réalisation de ce projet collaboratif à l'université Paul Sabatier Toulouse III, nous avons optimisé notre communication en créant un serveur Discord. Cette plateforme a facilité les échanges au sein du groupe, avec d'autres équipes, ainsi qu'avec les relais vers les intervenants externes.

En parallèle, pour un partage efficace des données et du travail, nous avons opté pour l'utilisation d'un Google Drive partagé. Cet espace centralisé contenait toutes les ressources essentielles, de la documentation relative au projet jusqu’aux rendus finaux. Le code a été réalisé au travers de notebooks Python, élaborés à l'aide de Google CoLaboratory, ce qui a permis un stockage direct sur Google Drive et une interaction simplifiée entre nos différents fichiers. Nous avons aussi créé un dépôt GitHub afin de permettre un développement collaboratif.

De plus, l'utilisation de la bibliothèque Gradio s'impose comme la meilleure solution de Data Visualisation possible. En effet, comme il s’agit d’un projet Open Data, nous tenions à utiliser une solution libre et accessible. Gradio offre une interface ergonomique pour la création de rapports, et nous voulions utiliser le langage Python, puisqu'il s’agit du langage de programmation avec lequel le groupe est le plus familier.

## Gestion du groupe

Dans le cadre de ce projet, notre objectif était de favoriser une cohésion de groupe forte et une implication active de chaque membre, tout en respectant des horaires de travail de 9h du matin à 17h le soir. Cette synergie a été cultivée à travers divers événements tel que le petit-déjeuner partagé, apporté chaque jour par un membre différent, permettant un petit moment de détente.

Chaque soir avant de partir, une séance de récapitulation permettait de faire le point sur les avancées et d'assigner les tâches du lendemain collectivement. Cette présentation des travaux réalisés par chaque membre a favorisé les discussions constructives et offert des perspectives pour l’évolution du projet. Lors de ces récapitulatifs, nous avons utilisé le projecteur afin d’illustrer nos propos. Ces moments, bien que paraissant initialement comme des pauses dans le travail effectif, ont prouvé leur efficacité en renforçant la compréhension globale du projet et en stimulant la créativité. Les échanges étaient fréquents pour clarifier des points spécifiques.

Pour garantir une compréhension approfondie du sujet, des variables et des modèles utilisés, une communication transparente a été maintenue. Les étudiants en deuxième année de Master ont joué à la fois un rôle de soutien, en aidant le reste du groupe grâce à leur expérience dans les domaines de la DataVisualisation et du Machine Learning et un rôle de développeur, en mettant en pratique cette expérience. La communication avec d'autres groupes a été organisée notamment pour les méthodes de création des Dashboard, tandis que les M2 ont créé des tutoriels afin de créer de la documentation plus précise et appropriée de l’utilisation de Gradio, en assurant ainsi une collaboration dynamique et productive.

## Problèmes rencontrés

Notre groupe a entamé ce projet avec de grandes ambitions: nous voulions pousser une analyse complète et approfondie de l’évolution des thématiques approchées par l’ensemble du paysage médiatique français. Cependant, en commençant le travail de valorisation des données, nous nous sommes rendus compte des limitations très fortes auxquelles nous étions exposés. Nous n’avions par exemple que très peu de données concernant le contenu des émissions télévisées/radio proposées. Les données à la source des visuels du dashboard concernant l’évolution des thèmes abordés lors des JT ne concernent que cinq chaînes, par exemple, et la granularité des thèmes permet très peu d’approfondissements.

Ces mêmes données devaient être utilisées pour regrouper les chaînes en fonction des thèmes abordés via une méthode de clustering, mais cette idée a été rapidement abandonnée pour cette même raison: regrouper 5 chaînes n’avait que peu d’intérêt. N’ayant pas pu trouver d’autres données pour compléter celles-ci, nous avons choisi d’explorer de nouvelles pistes, qui ont également amené leur lot de difficultés.

D’un côté, nous avons voulu compléter le travail effectué sur le temps de parole des femmes avec des données sur la représentations des femmes dans les films diffusés. C’était pour nous une façon de mettre en valeur les données sur les films dont nous disposions en les croisant avec des données externes sur la représentation des femmes comme par exemple le test de Bechdel. Nous avons vite trouvé un dataset avec ces informations mais le croisement a été complexe, notamment à cause des titres qui n’étaient pas dans la même langue d’un dataset à l’autre. Nous nous en sommes sortis en scrapant la traduction des titres sur wikipédia.

De l’autre côté, l’un des dataset a été assez complexe à aborder. Il s’agit d’une étude sous forme de questionnaire posé à un échantillon de 3000 français. Les réponses sont résumées dans un fichier excel de plus de 1000 colonnes et encodées par des nombres de différentes façons. La signification des colonnes et des codes a été fournie dans un autre fichier mais beaucoup de questions ont demandé un traitement spécifique, ce qui nous a freiné un certain temps. Ce dernier nous a permis d’ouvrir les sujets sur l’âge et la politique.

Nous avons passé énormément de temps à chercher des données ouvertes afin de compléter celles qui nous ont été mises à disposition, et ces recherches se sont souvent montrées peu fructueuses.

## Gantt du projet

Pendant les mois de novembre et décembre 2025, nous avons eu l'opportunité de participer à trois sessions animées par le Catalyseur portant sur la gestion de projet à travers la méthode du "Brown Paper". Ces séances nous ont permis d'anticiper les diverses tâches à accomplir, la répartition des charges de travail, ainsi que les différents risques face auxquels se prémunir tout au long du projet.

Ces sessions ont joué un rôle crucial en nous permettant non seulement de se projeter sur le déroulement global du projet et les éventuels problèmes qui pourraient survenir, mais elles ont également mis en lumière les grandes lignes des tâches à accomplir. Par conséquent, nous avons élaboré un diagramme de Gantt prévisionnel illustrant le déroulement général du projet, dans l'hypothèse d'une exécution sans problème particulier. Ci-dessous, vous trouverez un aperçu de ce diagramme de Gantt :

### *Figure 1 : Diagramme de GANTT prévisionnel*

Après avoir élaboré notre plan prévisionnel avec le diagramme de Gantt, nous sommes passés à la mise en action du projet. La réalité opérationnelle a souvent réservé des ajustements imprévus. Ainsi, jetons un coup d'œil sur notre Gantt réel pour voir comment le projet a réellement évolué. Cette comparaison nous permettra d'analyser les différences, de souligner les réussites et d'apprendre des leçons essentielles pour nos futurs projets.

### *Figure 2 : Diagramme de GANTT réel*

Comme on peut le constater, des différences se manifestent entre nos deux diagrammes de Gantt. Comme dit précédemment, de nouvelles sections du projet ont été créées en cours de route après avoir analysé nos données, le nombre de tâches est donc très différent. Nous avons pris du retard sur le traitement des données, notamment à cause du dataset “Enquête sur les français et l'information” dont la compréhension et le traitement ont été particulièrement longs et retors. La recherche et l’incorporation de données sur le cinéma, bien que non prévue dans le Gantt prévisionnel, a également été une source de retard à cause des problèmes rencontrés pour les ajouter à notre étude.

Au contraire, la plupart des autres datasets ont été simples à traiter, ce qui a permis de commencer plus tôt toutes les tâches de traitement (dashboard et clustering) initialement prévues. Le dashboard sur les thématiques abordées a été terminé très rapidement, ce qui a permis de fournir du support aux groupes en difficultés et de combler le retard.

# Description des données

Cette section propose une présentation approfondie du projet en explorant deux aspects cruciaux : le contexte qui éclaire les circonstances et les motivations autour des données, ainsi que leur contenu, leur structure et leur contenu essentiel, donnant ainsi les informations nécessaires pour la suite de notre analyse.

## Contexte

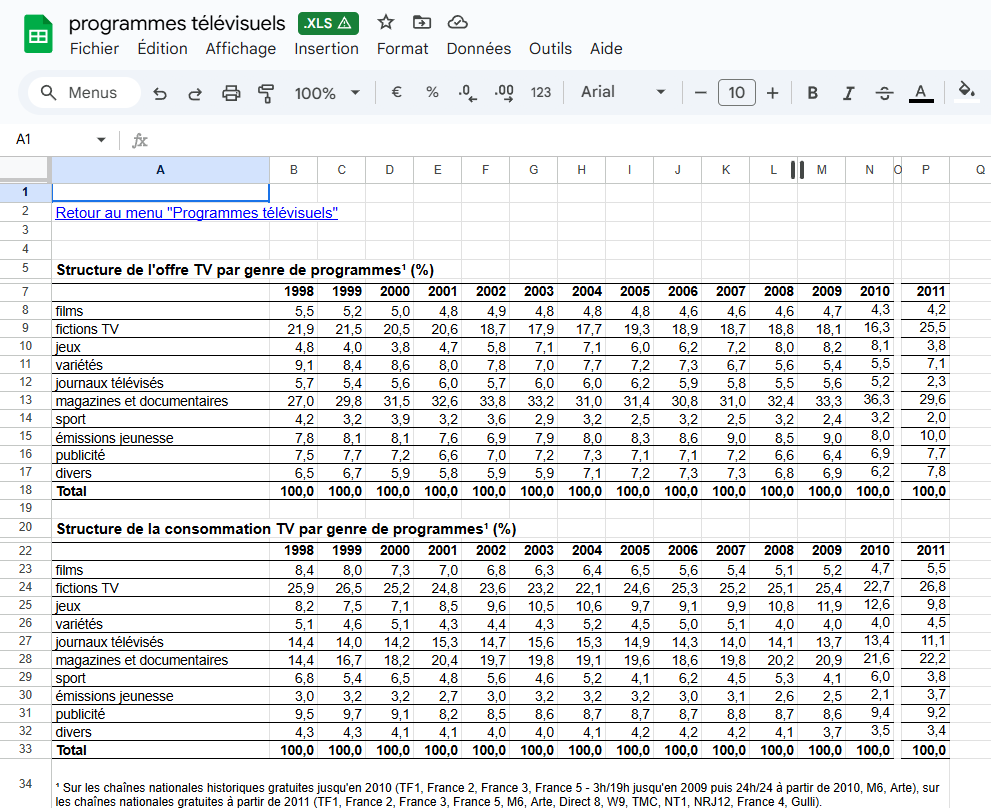
Les données se présentent sous forme de onze dépôts issus de trois sources publiques: l’INA (l’Institut National de l’Audiovisuel), l’Arcom (Autorité de régulation de la communication audiovisuelle et numérique) ainsi que du CNC (Centre national du cinéma et de l'image animée). Chaque dépôt comporte entre un et une dizaine de fichiers sous forme csv et xlsx principalement. Certains dépôts sont munis de documentation, tandis que d’autres sont simplement des tableaux sans en-têtes.

Les données de l’INA traitent des sujets des journaux télévisés ainsi que du temps de parole des femmes et des hommes à la télévision et à la radio et pour des programmes ayant fait l'objet d’une déclaration au CSA pour son rapport portant sur la représentation des femmes.

Les données de l’Arcom sont issues d’une étude sur les Français et l’Information, c’est le fichier qui a été le plus compliqué à traiter. Il contient beaucoup d’informations mais nous avons en particulier utilisé celles qui concernent les chaines regardées, les médias utilisés, et les opinions politiques.  
 Enfin, les données du CNC sont différentes statistiques concernant les chaînes françaises, telles que les audiences de la télévision, la consommation des ménages en vidéo à la demande, le classement des films les plus diffusés sur les chaînes nationales gratuites depuis 1957 ou bien le financement de la télévision.

## Exploration des données

En nous penchant sur les données qui nous ont été fournies pour ce challenge, nous nous sommes heurtés à plusieurs freins à la conception des dashboards. Premièrement, beaucoup de fichiers Excel nécessitent d’importants traitements, puisqu’ils présentaient plusieurs tableaux en une seule page, ou bien des données mal formatées ou contenant des valeurs incohérentes.   
  
 De plus, un des sondages de l’Arcom concernant la consommation de contenu audiovisuel par les français semble être très riche en informations, cependant le format des réponses à ce questionnaire est assez opaque, puisqu’il y a de très nombreuses variables et de nombreux formats de questions/réponses.



### Figure 3 : Exemple de contenu d’un fichier csv

Enfin, pour certains sujets les données sont très limitées, comme par exemple pour le clustering des chaînes par thèmes abordés, nous n’avons que des données que sur cinq chaînes, ce qui réduit considérablement l’utilité du clustering.

# Méthodologie

Cette partie constitue le pilier de notre approche dans la réalisation de ce projet. Nous examinerons en détail les différentes étapes et techniques que nous avons adoptées pour accomplir nos objectifs. De la gestion des données à l'application des algorithmes, en passant par l'analyse statistique, cette partie offre un regard approfondi sur notre méthodologie. Chaque décision a été prise en soulignant nos partis pris afin d'assurer la fiabilité et la pertinence de nos résultats dans l'analyse de la dangerosité des bâtiments.

Au début de notre projet, lors de la phase d’organisation et de répartition des tâches organisée avec le catalyseur, nous avions décidé de mettre en place des algorithmes de clustering afin d’identifier des chaînes proposant des contenus semblables. Initialement, nous pensions avoir des dataset très complets, traitant non seulement des thèmes mais aussi des contenus des émissions. Cependant, comme mentionné précédemment, nous n’avons eu qu’un dataset comportant six colonnes et traitant de cinq chaînes différentes. Nous n’avions pas beaucoup de paramètres sur lesquels effectuer du clustering, et de toute façon, il semble peu pertinent d’effectuer du clustering pour uniquement cinq chaînes. Nous nous sommes donc contentés du rapprochement des chaînes via la cos-similarité des répartitions moyennes des temps d’émissions par thèmes vue dans le premier dashboard.

## Méthodologie de création des tableaux de bord

La création des dashboard est une mission très délicate puisqu’elle demande de choisir judicieusement les filtres et visuels afin de valoriser les données un maximum. Une des erreurs que nous avons souhaité éviter était de se lancer directement dans du code sans réellement savoir quelle serait l’organisation finale du tableau de bord. Pour éviter ce problème, nous avons mis en place un processus de création de dashboard, prenant forme de trois étapes:  
  
- La création de sketchings: une maquette sur papier du tableau de bord, afin d’identifier les meilleurs visuels et de déterminer le meilleur ordonnancement de ces derniers.

- La création des visuels en python, séparée parmi les différents membres de l’équipe.  
  
- Le regroupement de l’ensemble de ces visuels en un dashboard, en suivant la maquette définie au préalable.

## Premier tableau de bord: Thématiques

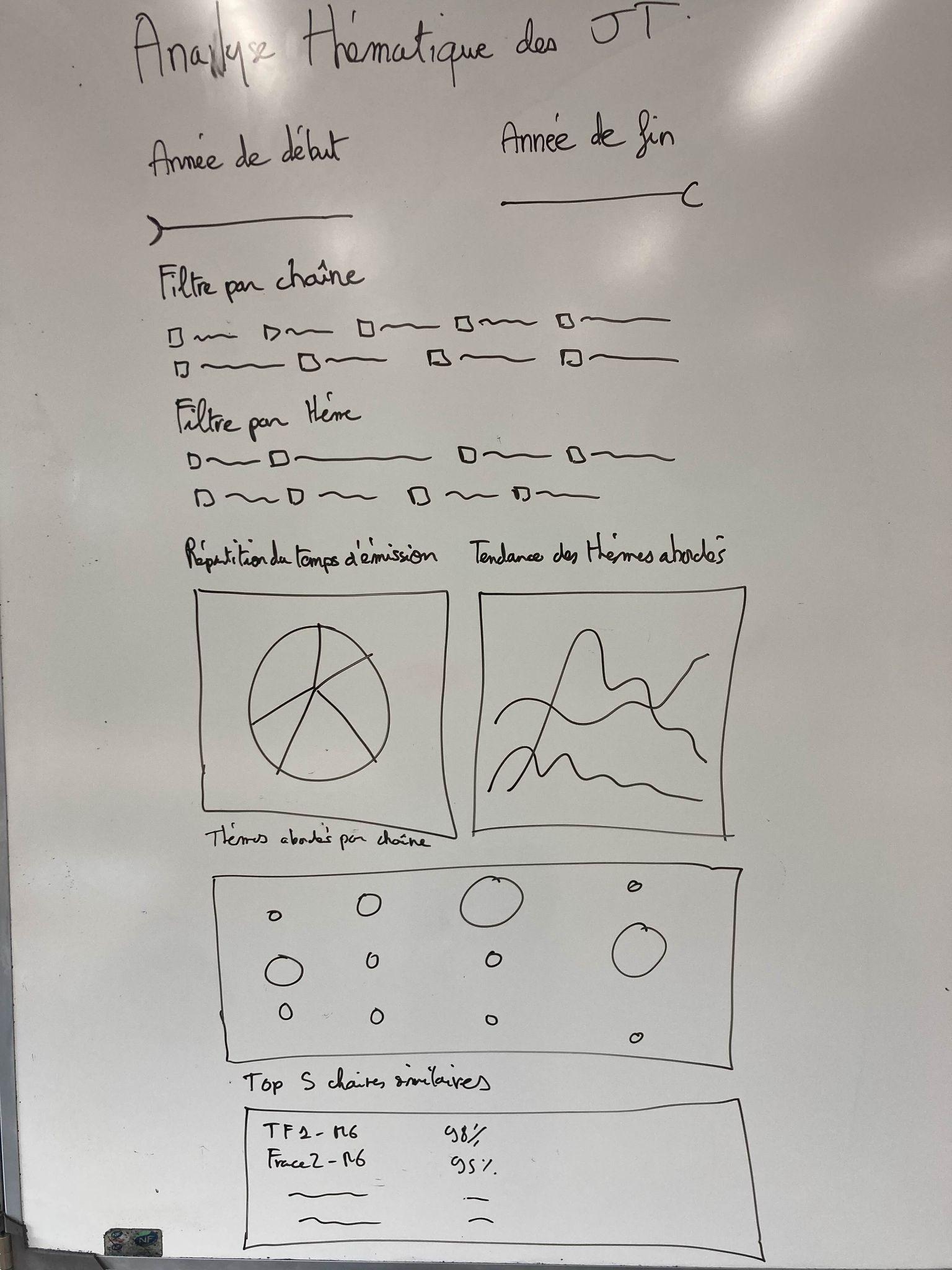
Le premier tableau de bord se base sur un baromètre de l’INA qui présente le nombre de thèmes abordés quotidiennement par cinq chaînes françaises ainsi que leur durée. Ces données s’étalent de janvier 2000 au 31 décembre 2020. On a donc deux décennies de données prêtes à être analysées, ce qui correspond à un dataset de près de 280 000 lignes.

Les données comportent cinq colonnes:

* La date de l’émission
* La chaîne concernée
* Le thème concerné
* Le nombre d’émissions concernant ce thème
* Le temps d’émission ce jour

Les données ont été acquises via un algorithme développé par l’INA, qui détecte automatiquement les thèmes abordés au cours des JT.  
Un des problèmes que nous avons soulevés est que pour une part assez importante d'émissions, la durée d’émission renseignée par l’algorithme (et donc présentes dans nos données) est de seulement quelques secondes. Par exemple, on a environ cinq cent émissions de durée d’une seconde.

Il s’agit de valeurs aberrantes, et la question a été de savoir s’il fallait les supprimer.  
Après réflexion, nous avons décidé de ne pas le faire pour plusieurs raisons. Premièrement, nous n'exploitons pas le nombre d’émissions concernant les différents thèmes, donc avoir un nombre d’émissions trop important n’importe pas. De plus, ces émissions superflues n’ajoutent que quelques secondes par an d’émission par thème, ce qui est clairement négligeable. Enfin, supprimer des émissions d’une durée courte nécessite de déterminer une limite minimum de temps d’émission, ce qui est assez compliqué à prendre en considération, puisque cette limite diffère en fonction des thèmes, des types d’émissions (intervention spéciale, émission sans reportage vidéo, interruption ou annonce…).  
  
 Après s’être familiarisé avec les données, nous avons pu nous lancer dans la conception de maquettes en équipe au tableau.



### Figure 4 : Maquette du premier dashboard

Le premier visuel de ce dashboard est un diagramme circulaire permettant de montrer la répartition des temps d’émission en fonction des chaînes. Cela permet de se rendre compte des disparités de représentativités de certaines données.

Le deuxième visuel montre l’évolution des thèmes abordés en fonction du temps, pour l’ensemble des chaînes sélectionnées. Ce visuel a pour objectif de capturer des tendances à la hausse ou à la baisse de certains thèmes, ou bien des saisonnalités de certains d’autres d’entre eux. On s’attend par exemple à voir une saisonnalité du thème “Sports” tous les quatre ans pour les jeux Olympiques d’été, qui est un événement largement couvert en France.

Par la suite, on a opté pour un diagramme en bulles pour représenter la part d’émission allouée en moyenne à chaque thème par chaîne. C’est un moyen très visuel et très pratique pour représenter de nombreuses informations, et cela permet de comparer d’un coup d'œil la part globale de temps d’émission que prend chaque thème.

Enfin, le dernier Visuel de ce tableau de bord est un top 5 des chaînes similaires, en se basant sur la part de temps d’émission allouée à chaque thème. Pour mesurer la similarité entre ces chaînes, on s’est basé sur une similarité cosinus. L’objectif est de détecter comme proches deux chaînes qui allouent le plus de temps aux mêmes sujets.

Ces quatre visuels permettent d’exploiter au maximum les données présentes dans le fichier, il est important de souligner que l’on n’a que 5 colonnes dans les données.



*Figure 5 : Rendu final du premier dashboard*

Une fois la phase de développement terminée, voici le rendu final du dashboard.

## Second tableau de bord: Temps de parole des femmes

Ce second tableau de bord s’appuie sur plusieurs fichiers de données fournis par l’INA.

Le premier que nous avons utilisé (20190308-years.csv) recense le taux d’expression des femmes par chaîne et par année, ainsi que le taux global de parole. Les données couvrent la période de 1995 à 2019 pour la radio, et de 2010 à 2019 pour la télévision. Au total, 55 chaînes de télévision et stations de radio ont été analysées, représentant un volume de 1 075 978 heures de programmes étudiés.

Le dataset contient 7 colonnes :

* Le type de média (tv ou radio)
* Le nom de la chaîne/station
* Si c’est une chaîne publique ou non
* L’année étudiée
* Le taux d’expression des femmes
* Le taux de parole
* Le nombre d’heures analysées

Comme pour les données utilisées avec le premier dashboard, le taux d’expression des femmes ainsi que le taux de parole ont été obtenues grâce au logiciel libre “inaSpeechSegmenter” développé par l’INA. Il détecte les zones de musique et les zones de parole contenues dans les documents audiovisuels.

Pour nous différencier de l’article de l’INA et au vu du nombre de chaînes nous avons décidé de les regrouper par thèmes. L’objectif fut donc d’ajouter pour chacun des csv que nous allions utiliser, une colonne ‘Thème de la chaîne’. Voici les thèmes que nous avons sélectionnés : Musique, Information, Culture, Divertissement, Sport et Tout. Ces thèmes ont été choisis par nous même en nous renseignant sur chaque média et son contenu. Ainsi nous avons conçu un dictionnaire qui associe à chaque thème, les chaînes correspondantes et qui pourra être utilisé pour enrichir aisément chaque csv de données.

L’objectif de ce Dashboard est de montrer les inégalités de représentation des femmes à la télévision et à la radio.

Le premier graphique sert d'introduction au sujet en présentant la répartition du temps de parole entre les femmes et les hommes sur l'ensemble des médias, année par année.

Ensuite nous avons fait un diagramme montrant l’évolution du taux de paroles des femmes au fil des années selon les thèmes.

Puis nous avons représenté par un histogramme la différence entre le taux d’expression des femmes en 2019 et en 2010 (ou en 1995 pour les stations de radio) pour chaque chaîne et station. Cette visualisation permet de dresser un bilan de l’évolution du temps de parole des femmes pour chacune d’elles.

Après avoir vu l’évolution des temps de parole au fil des années, nous avons souhaité utiliser un autre jeu de données mise à disposition par l’INA intitulée *hoursatall.csv* qui pour un média (tv ou radio), une année et une tranche horaire de la journée associe un taux de parole des femmes.

Ce fichier csv contient donc les colonnes du fichier *years.csv* qui ont été décrites précédemment ainsi que l’ajout d’une colonne ‘hour’ qui représente la tranche horaire. A noter que les tranches horaires s’étalent de 5h à 23h et que les données avec un nombre d’heures analysées inférieur à 20 sont exclues de la table.

Ainsi, il est important de préciser que naturellement la précision du taux d’expression des femmes dans ce fichier reste naturellement inférieure à celles du fichier *years.csv* qui ne distingue pas les tranches horaires.

A l’instar du fichier years les données s’étalent de 1995 à 2019 pour la radio et de 2010 à 2019 pour la télévision.

Avec ce jeu de données, la possibilité de construire un graphique montrant le taux d’expression des femmes dans les médias audiovisuels en fonction des tranches horaires nous a semblé pertinente. En effet, sachant que les heures de forte audience à la radio sont le matin et en fin d’après-midi/soir et que les heures de forte audience de la télévision sont le soir, il peut être intéressant de voir si l’on constate une nette différence dans le taux d’expression des femmes entre les pics d’audience et le reste du temps.

De tels graphiques sont proposés dans l’article de l’INA « A la radio et à la télé, les femmes parlent deux fois moins que les hommes ». Dans cet article, un graphique montre pour chaque média radio le pourcentage de taux de parole des femmes en fonction des tranches horaires ainsi que la moyenne publique et privée. De même, un graphique similaire pour les médias tv a été dressé. Le premier objectif a été de reconstruire ces graphiques mais en optant pour des options de visualisations différentes. En effet, afin d’essayer de visualiser ces graphiques en fonction des tranches horaires, nous avons choisi de regrouper les médias suivant les thèmes ou la catégorie du média (privé ou public). Ainsi nous obtenons des courbes non pas pour toutes les chaînes mais pour des critères qui seront les filtres de notre dashboard (thème de la chaîne, média privé ou public et nature du média).

Nous espérions pouvoir constater des tendances pour certaines catégories de médias dans l’évolution du pourcentage de taux de parole.

Ensuite, A l’aide du dataset *years.csv* décrit précédemment, nous avons voulu construire le graphique du taux d’expression des femmes en fonction du pourcentage de parole globale du média (100 - pourcentage de musique). Ainsi, ce graphique est sous la forme d’un nuage de point où chaque point correspond à un média (tv ou radio). Les points sont coloriés suivant le thème de la chaîne. L’objectif est encore une fois de voir si le taux de parole des femmes peut être corrélé avec le pourcentage de parole global et le thème de la chaîne. Ce graphique est naturellement dynamique car il dépend des filtres choisis.

Ensuite, nous avons réalisé une étude sur les médias incluant les programmes ayant fait l’objet d’une déclaration au CSA pour ne pas avoir assez donné du temps de parole aux femmes. Cette étude repose sur les données du fichier des données *ina-csa-parole-femmes-chaines.csv*, fournit par l’INA, contenant des déclarations de 2019 et 2020.

Ce jeu de données comprend, pour chaque éditeur (nom de chaîne ou de radio):

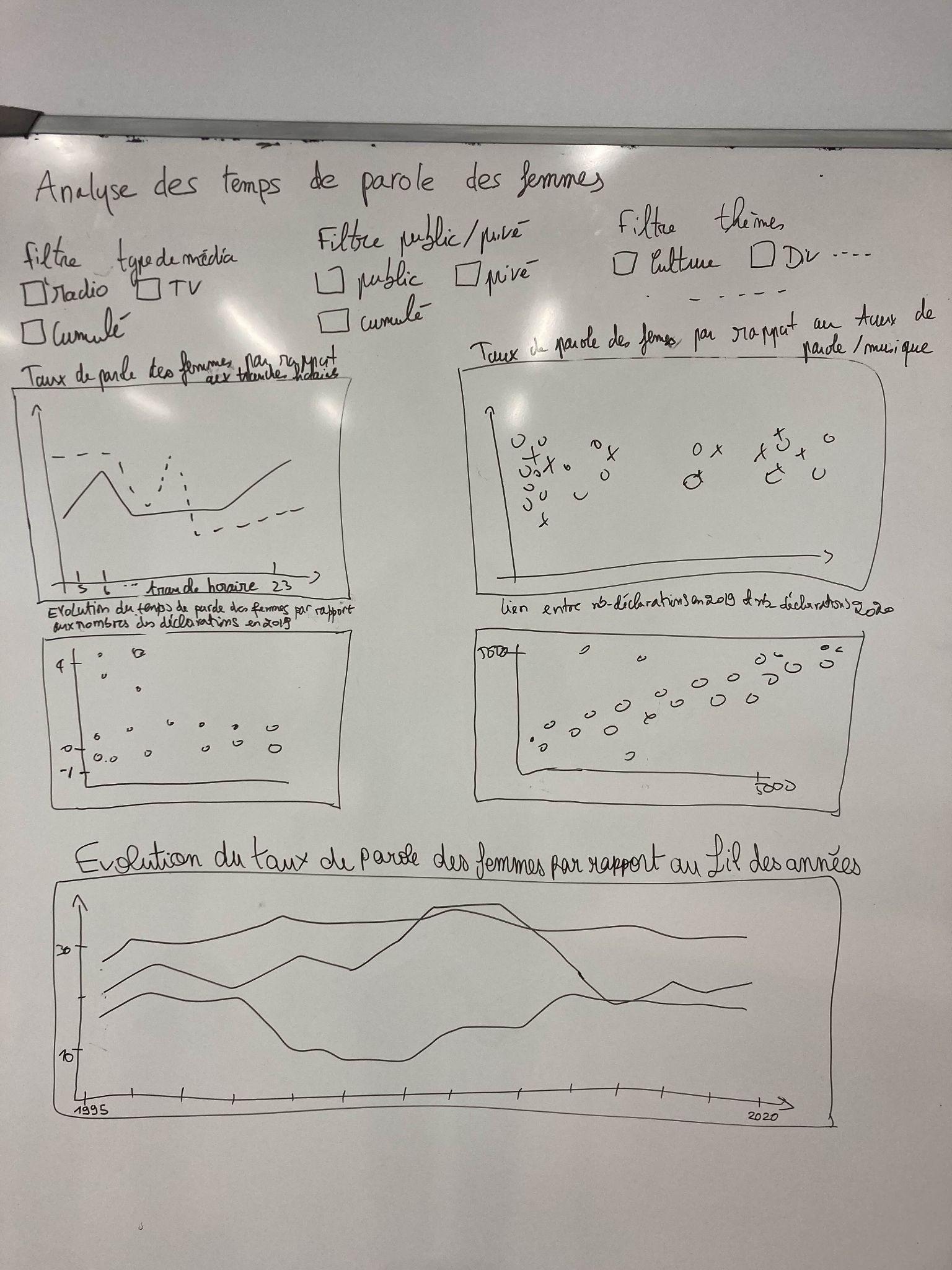
* le groupe auquel appartient l’éditeur
* Nombre de programmes ayant fait l’objet d’une déclaration en 2019 et en 2020
* Durée cumulée, en secondes, des programmes déclarés en 2019 et en 2020
* Durée cumulée, en secondes, de parole de femmes en 2019 et en 2020
* Durée cumulée, en secondes, de parole d’hommes en 2019 et en 2020
* Durée cumulée, en secondes, de musique, bruit ou silence en 2019 et en 2020
* Taux d’expressivité des femmes en 2019 et en 2020 (part des femmes dans le temps de parole total)
* Taux de parole par rapport au temps total en 2019 et en 2020

A l’instar des fichiers de données décrits précédemment, nous avons enrichi les données initiales en ajoutant des informations sur le caractère public ou privé des médias, ainsi que les thèmes généralement abordés par ces derniers.

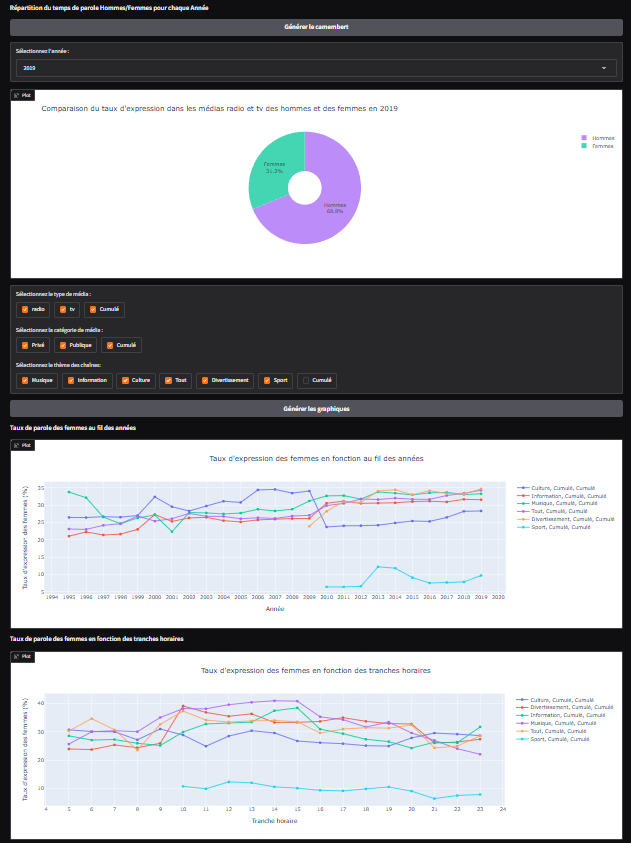
Pour mener l’étude, tout d’abord, nous avons commencé à explorer les données fournies par L’INA avec le fichier de données *ina-csa-parole-femmes-chaines.csv***.** Ce dernier recense 40 médias ayant des programmes déclarés en 2019 et 2020, répartis comme suit : 25 médias privés et 15 publics, comprenant 26 chaînes de télévision et 14 stations de radio, regroupés dans 12 groupes de médias. Parmi eux, le groupe M6 compte 7 médias ayant déclaré des programmes. En moyenne, le nombre de déclarations par média est passé de **1139,4** en 2019 à **1383,9** en 2020, soit une augmentation de plus de **24%**. Par ailleurs, le pourcentage moyen de temps de parole des femmes a progressé de **33%** en 2019 à **36%** en 2020, soit une hausse de **3%**. Ensuite, nous avons visualisé, l’évolution du taux d’expressivité des femmes entre 2019 et 2020 en fonction du nombre de programmes ayant fait l’objet d’une déclaration en 2019. De plus, nous avons étudié l’influence des déclarations des programmes des médias en analysant la relation entre le nombre de programmes déclarés en 2019 et en 2020.

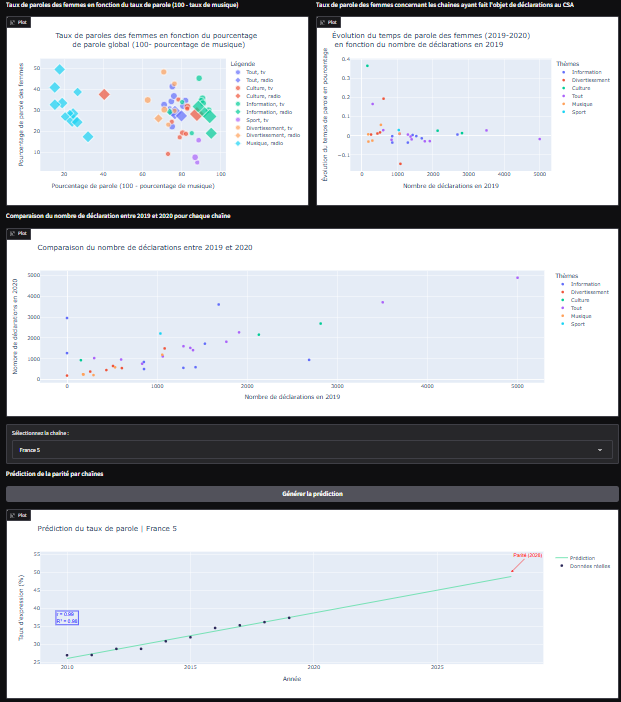
En dernier lieu, nous avons développé un graphique dynamique permettant de prédire l’année à laquelle la parité pourrait être atteinte suivant chaque média télé ou radio si l’on conserve la tendance actuelle.

L’idée était d’avoir une représentation visuelle de la tendance de l’évolution du taux d’expression des femmes à partir de l’ensemble des données recueillies pour chaque chaîne. Pour ce faire, quand une chaîne est sélectionnée, le coefficient de corrélation linéaire entre les années et le taux d’expression des femmes ainsi que le R² est calculé. Si la valeur absolue de ce dernier est supérieure à 0.5, nous considérons alors la modélisation de la régression linéaire afin de prédire la tendance du taux d’expression des femmes pour les prochaines années. Évidemment, cette approche reste assez primaire car elle ne considère uniquement des modèles de régression linéaire.



*Figure 6 : Plan du premier dashboard*





*Figure 7 : Rendu final du premier dashboard*

## 

## Complément: la représentation des femmes dans les film

### Représentation des femmes parmi les films avec le plus d’audience sur les chaînes françaises

Comme nous disposions de données sur les films les plus vus et diffusés à la télévision française, nous avons souhaité compléter les conclusions sur le temps de parole des femmes par des données externes sur leur représentation dans les films.

Pour mener à bien la création de ce tableau de bord, nous avons donc utilisé 2 dataset, “Meilleur audience.xlsx” et “TMDb\_Dataset.csv”. Le premier dataset, après une mise en forme, contient les films ayant fait le plus d’audience sur une certaine chaîne à une certaine année, on notera aussi qu’avec chaque film il y avait le nombre de téléspectateur et la part d’audience, c’est à dire la part que représente l’émission par rapport à toute les autres. Le second dataset, “TMDb\_Dataset.csv”, vient des données d’IMDB (Internet Movie Database). Ce jeu de données recense une grande partie des films qui existent en y faisant correspondre, pour ce qui nous intéresse, l'âge de l’acteur et de l’actrice principale (l’acteur ou l’actrice la plus importante dans le film) et le test de bechdel.

Pour expliquer simplement ce qu’est le test de bechdel, c’est une note allant de 0 à 3. Si le film possède au moins 2 actrices nommées il a 1 point, si il y a une conversation entre elles le film a 2 point et si cette conversation porte sur autre chose que d’un homme alors le film a 3 point, sinon il a 0. Ce test nous permet d’évaluer la représentation des femmes dans les films.

Ce tableau de bord a pour but d’essayer de voir des différence de répartition entre les chaînes des notes aux tests de bechdel des films qui ont fait le plus d’audience pour ces chaînes et la différence des différence d’âge entre l’actrice et l’acteur principale entre les chaînes et entre les notes bechdel.

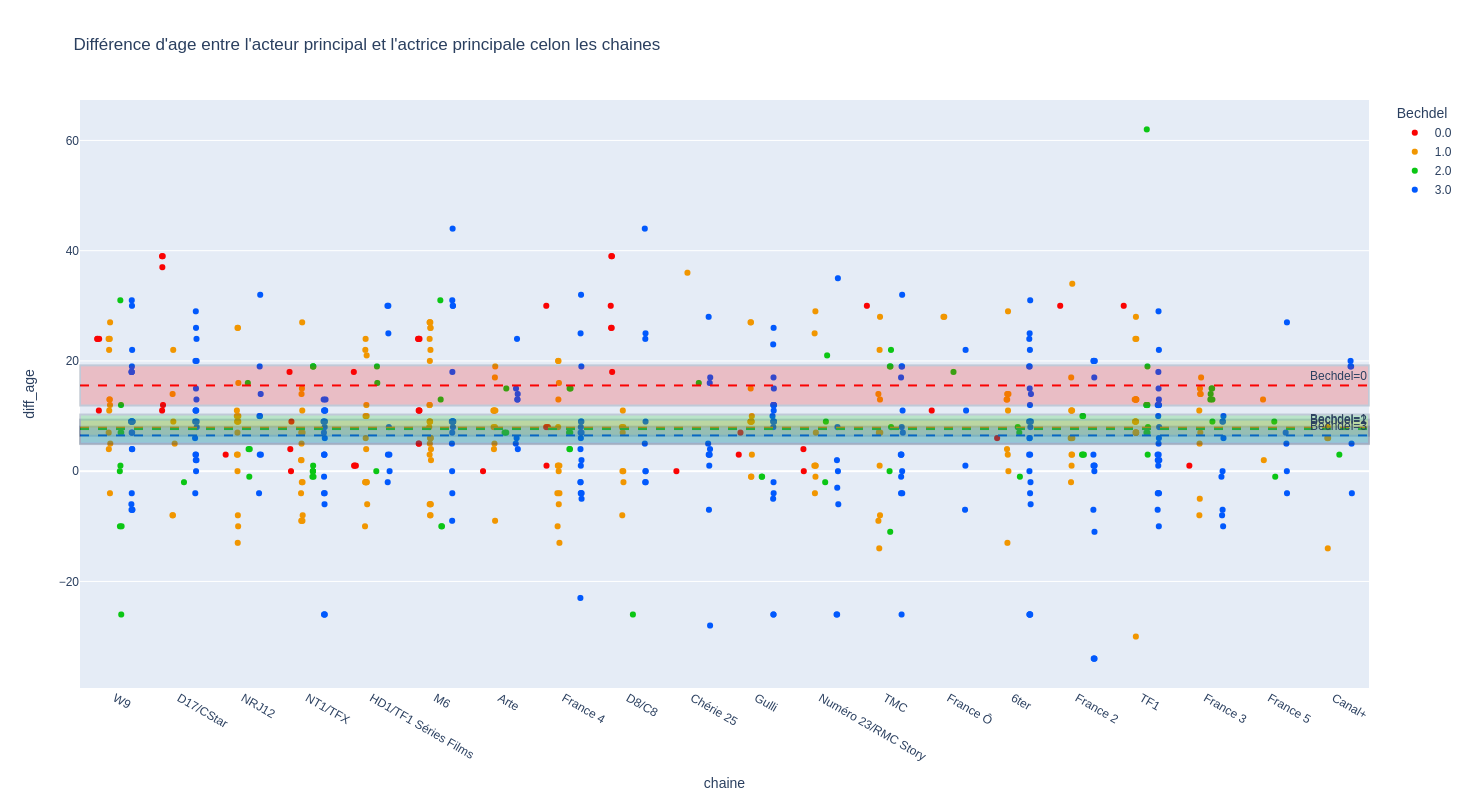
L’affichage peut être filtré par période et par chaîne. Les chaînes changent parfois de noms, donc nous avons regroupé ces noms sous un seul de cette forme “ancien\_nom/nouveau\_nom”.

Comme les données ne sont pas toujours suffisantes pour avoir une bonne représentation, on peut, au lieu de regarder par chaîne, regarder par groupe.

Par souci de temps il n’y a pas d’option groupe de chaîne mais voici la liste de ces groupes :

* Groupe TF1 : TF1, TMC, NT1/TFX, HD1/TF1 Séries Films
* Groupe M6 : M6, W9, 6ter, Guilli
* France Télévision : France 2, France 3, France 4, France 5, France Ô
* Groupe Canal+ : Canal+, D8/C8, D17/CStar
* NRJ Group : NRJ 12, Chérie 25
* Arte et Numéro23/RMC Story n’ont pas de groupe pour notre visualisation

Le graphique ci-dessous affiche pour chaque chaîne les points représentant les film selon la différence d’âge entre l’actrice et l’acteur principale, tout en les groupant par leurs notes au test de bechdel. De plus il y a une représentation de la moyenne de la différence d’âge par note aux test avec leurs intervalles de confiance, par exemple la ligne rouge montre la moyenne de des différence d’âges pour les films ayant une note de 0 au test de Bechdel.



*Figure 8 : Tableau de bord sur la différence d’âge*

### Représentation des femmes parmi les films les plus diffusés à la télévision française depuis 1957

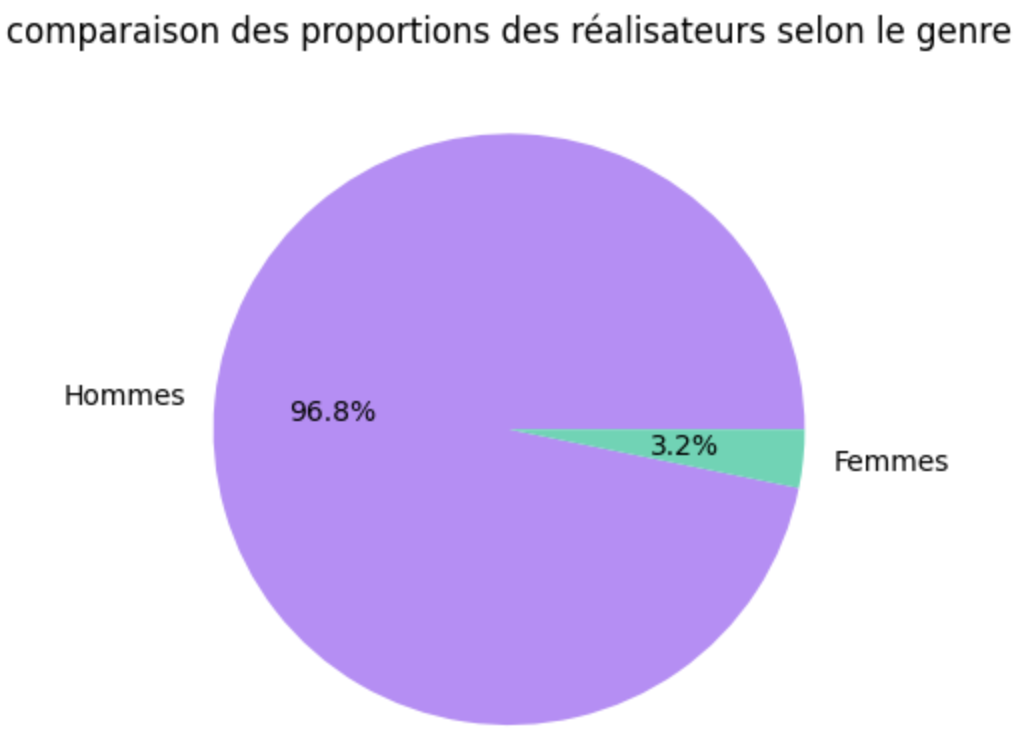
L’objectif de cette étude est d’analyser la représentation des femmes dans les films les plus diffusés sur les télévisions françaises depuis 1957. Cette analyse repose sur le fichier de données **LucasFeatAlain.csv**, élaboré à partir du fichier **Films les plus diffusés à la télévision.xlsx**, fourni par le Centre national du cinéma et de l'image animée (CNC), ainsi que sur des informations issues de la base de données IMDb (Internet Movie Database).

Le fichier initial contient le classement de 258 films les plus diffusés depuis 1957. Cependant, en raison de problèmes d’extraction depuis IMDb, nous avons pu récupérer les informations détaillées (réalisateurs, acteurs, etc.) pour **189 films seulement**.

Le fichier **LucasFeatAlain.csv** contient ainsi les données suivantes pour chaque film :

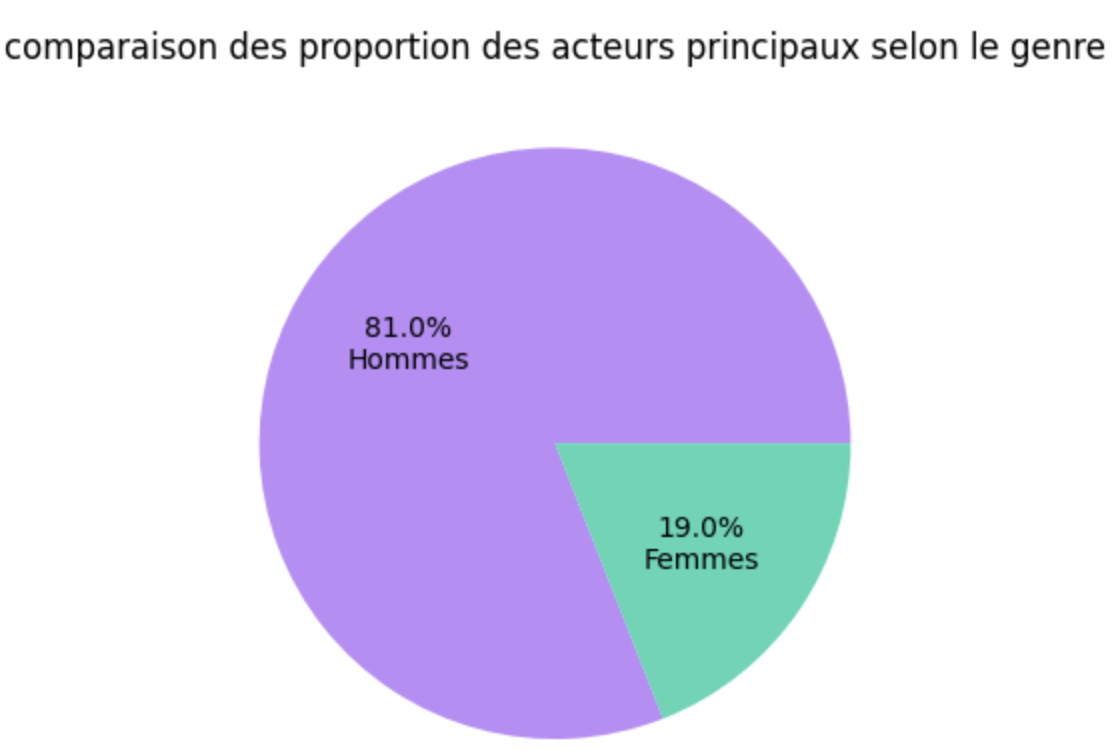
* Le titre ;
* Le réalisateur ;
* Les dates de production et de sortie ;
* Le nombre total de diffusions depuis 1957 ;
* Le nombre moyen de diffusions par an ;
* Une indication sur le genre du réalisateur (homme ou femme) ;
* Une indication sur le genre de l’acteur principal (homme ou femme) ;
* Le pourcentage d’actrices parmi le casting total.

Pour faire cette analyse, nous avons commencé d’abord à comparer les proportions des réalisateurs et des acteurs selon le genre :



*Figure 9 :Genre des réalisateur.ices de films*

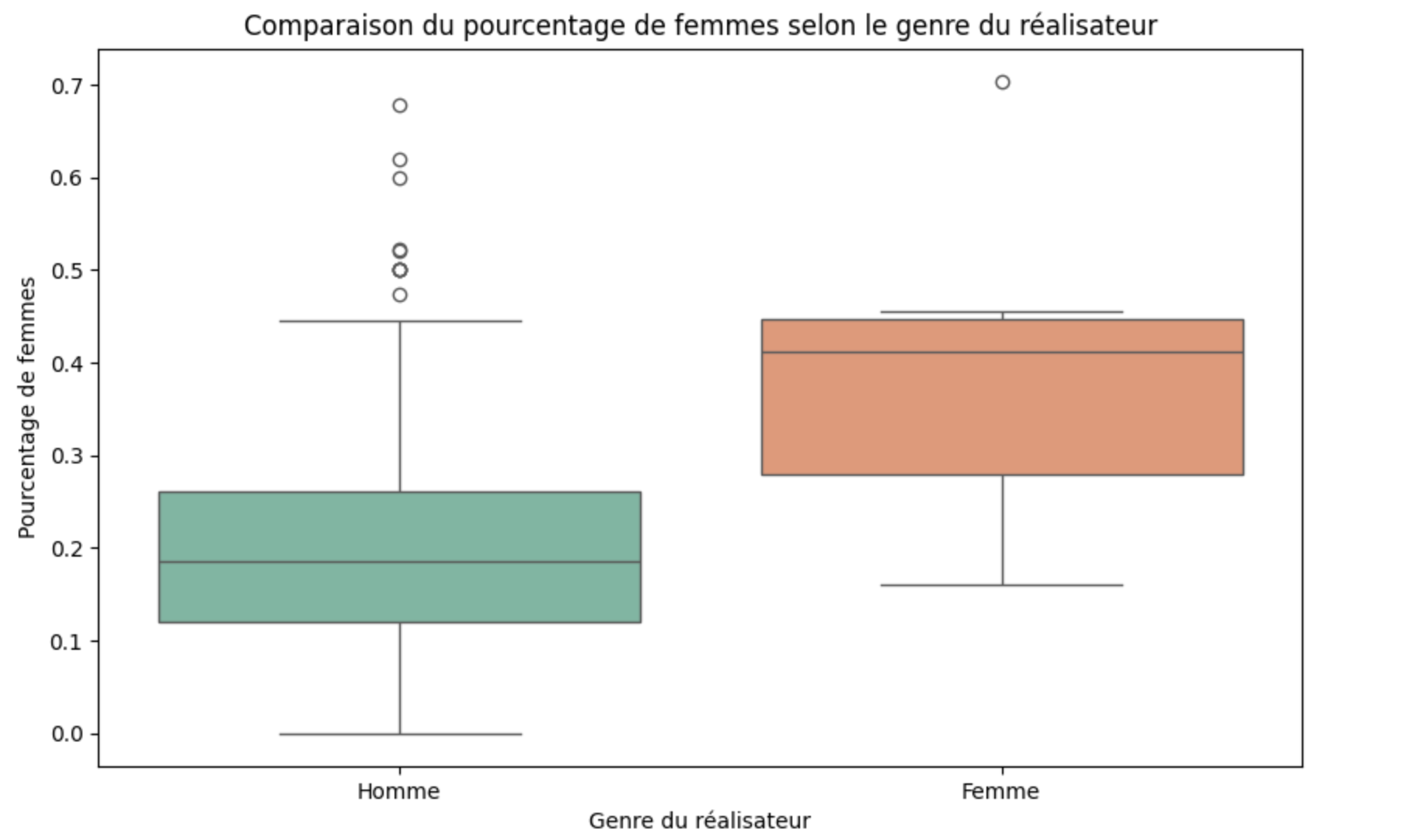
Ce graphique illustre la proportion de femmes réalisatrices parmi les films de notre jeu de données. Il met en évidence qu'elles ne représentent que 3,2 % du total, contre 96,8 % pour les hommes.



*Figure 10 : Genre des acteur.ices de films*

En ce qui concerne les acteurs principaux, la proportion d'actrices principales est meilleure que celle des réalisatrices, mais reste encore assez faible : seulement 19 % des rôles principaux sont tenus par des femmes, contre 81 % par des hommes.

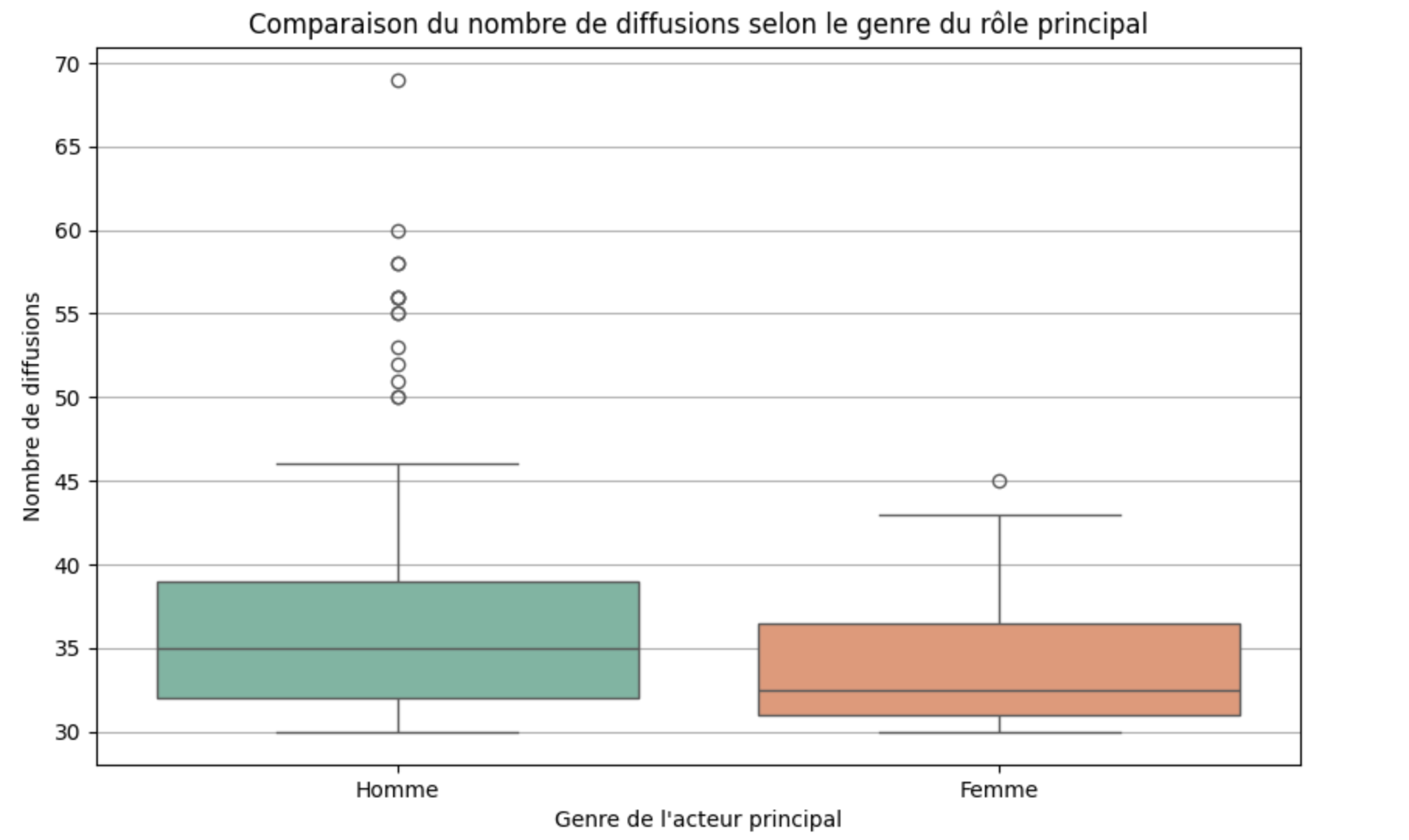
Ensuite, nous avons comparé le pourcentage des femmes parmi les acteurs des films selon le genre du réalisateur



*Figure 11 : Part de chaque genre d’acteur.ices par genre du réalisateur.ice*

Ce graphique nous montre la distribution du taux des femmes actrices selon le genre du réalisateur. On remarque que la médiane du pourcentage de temps de parole des femmes est plus élevée pour les réalisatrices (autour de **40 %**) que pour les réalisateurs (environ **20 %**), ensuite, on réalise que pour les réalisateurs les pourcentages des femmes actrices sont très faibles (jusqu’à 0%) et rarement au-dessus de 50 % tandis que pour les réalisatrices on a au moins environ 15% des femmes. Cela suggère que les réalisatrices donnent, en général, une place plus importante aux femmes dans leurs films.

Enfin, nous avons analysé le nombre de diffusions des films selon le genre des acteurs principaux



*Figure 12 : Nombre de diffusions par genre du rôle principal*

Ce graphique nous montre la distribution du nombre de diffusions des films selon le genre de l’acteur principal. On remarque que, généralement, les films dont une femme est l’actrice principale sont moins diffusés que ceux dont l’homme est l’acteur principal avec une médiane égale à environ 20 pour chez les femmes et 35 chez les hommes. On remarque aussi une plus grosse variabilité chez les hommes que chez les femmes.

​​Cette visualisation met en lumière des inégalités persistantes dans la représentation des femmes, aussi bien en tant que réalisatrices qu’en tant qu' actrices principales dans les films les plus diffusés à la télévision française depuis 1957. Les femmes réalisatrices restent largement sous-représentées, et les films avec des actrices principales bénéficient d’une diffusion moindre.

## Les films diffusés à la TV

Le but de cette partie est d’analyser les données de diffusion des films sur plusieurs chaînes de télévision afin d’identifier les tendances, la répartition des films français et américains, ainsi que leur contribution globale. Puis nous Nous avons procédé en plusieurs étapes : nettoyage des données, division en sous-groupes (films français, américains) et visualisation des résultats sous différentes formes.

#### **Nettoyage des Données:**

* **Description des données initiales** : Le jeu de données contient des informations sur les films diffusés, avec les chaînes de télévision en lignes et les années en colonnes. Chaque cellule représente le nombre de films diffusés pour une chaîne donnée sur une année précise.
* **Problèmes rencontrés** :
  + Valeurs manquantes ou incorrectes (ex :"-").
  + Présence de lignes inutiles.
* **Actions de nettoyage** :
  + Remplacement des tirets "-" par 0.
  + Suppression des lignes contenant des valeurs manquantes (nd).

Après le nettoyage, les données ont les divisées en trois parties :

* **Films français**: Extraction des données pour les films français uniquement.
* **Films américains**: Extraction des données pour les films américains uniquement.
* **Nombre total de films (Fr + Am)** : Fusion des deux parties précédentes pour une vue globale.

#### **Visualisations et Analyses:**

Ce tableau de bord interactif permet d'explorer les données de diffusion des films français et américains sur différentes chaînes de télévision entre 1990 et 2023.



*Figure 13 : Rendu final du tableau de bord sur les films*

**Évolution des données par chaîne (Line Chart)** :

Ce graphique représente l'évolution des nombres de films diffusés pour différentes chaînes de télévision sur une période allant de 1990 à 2023. Chaque chaîne est représentée par une courbe d'une couleur spécifique. TF1 domine largement les autres chaînes avec des valeurs atteignant presque 1000 avant de diminuer après 2010. France 2 a des valeurs stables autour de 400 jusqu’en 2010, puis une baisse progressive. France 3 reste stable, autour de 200-300, mais décline légèrement après 2010. France 4 reste dans des valeurs basses. Canal+ connaît une certaine stabilité mais reste en deçà des chaînes publiques. C8, W9, TMC: Les chaînes apparues après les années 2000 montrent une croissance modeste et se stabilisent autour de 100-200. Les courbes pour les "sous-total chaînes nationales publiques" et "sous-total chaînes nationales payantes" agrègent les données des groupes correspondants, permettant une vue d'ensemble. Observations générales : TF1 reste la chaîne dominante jusqu'à un déclin marqué après 2015. Les chaînes publiques comme France 2 et France 3 connaissent une baisse d’audience après 2010. Les chaînes plus récentes, comme C8, W9, et les autres, prennent une part de marché croissante, sans dépasser les chaînes historiques. Ce graphique illustre donc bien les changements dans la consommation télévisuelle en France sur 30 ans, avec une montée en puissance des chaînes privées et une diversification de l'offre, mais ce qu'on remarque c'est qu'après 2020 la consommation télévisuelle a diminué.

**Répartition des films diffusés par chaîne (Camembert) :**

En résumé, la catégorie "Autres" regroupe des chaînes qui, individuellement, n'ont pas un impact majeur, c’est les chaînes avec des parts inférieures à 5 %, mais collectivement, elles jouent un rôle crucial en couvrant environ un tiers de la répartition des films diffusés. Cela met en avant l'importance de ces chaînes "minoritaires" dans l'offre cinématographique globale. Sous-total des chaînes publiques (23,8 %)

Les chaînes publiques (comme France Télévisions) jouent un rôle important dans la diffusion de films depuis leur création, ce qui reflète leur mission culturelle et éducative. Sous-total des chaînes payantes (13,3 %)

Ce pourcentage, incluant des chaînes comme Canal+, montre leur importance dans l'industrie du cinéma depuis leur lancement. Canal+ est particulièrement connu pour ses accords avec l'industrie cinématographique française et internationale. Canal+ (13,3 %)

Canal+, créée en 1984, a été un acteur majeur dans la diffusion de films grâce à ses partenariats avec des studios de cinéma, garantissant des exclusivités sur des films récents. Arte (9,9 %)

Créée en 1992, Arte se concentre sur le cinéma d'auteur et les films indépendants, ce qui explique sa part importante malgré son positionnement de niche. France 3 (6,2 %)

En tant que chaîne publique existant depuis 1975, France 3 a contribué à la diffusion de films, mais sa programmation se concentre davantage sur des contenus locaux et des séries. TF1 (5,2 %)

Première chaîne privée en France (lancée en 1975), TF1 diffuse des films, mais cette activité est moins centrale par rapport à d'autres types de programmes (divertissements et séries).

**Distribution des pourcentages (Boxplot) :**

Sous-total chaînes nationales publiques : La boîte est large, avec des pourcentages souvent élevés. Elle atteint un maximum important proche de 25 %, ce qui montre une forte contribution dans les films diffusés.

Sous-total chaînes nationales payantes : ces chaînes sont également importantes dans la diffusion de films. La distribution est large, et les pourcentages sont souvent supérieurs à ceux des chaînes individuelles.

France 3 et France 2 : Les pourcentages sont significatifs, centrés entre 5 % et 15 %. Cela reflète leur rôle majeur dans la diffusion cinématographique.

Canal+ : Une chaîne très importante pour les films, avec une distribution étendue et une médiane élevée (autour de 10 %).

Arte : Sa contribution est modeste (autour de 5 %), ce qui est cohérent avec son focus sur les films indépendants et d’auteur.

TF1 et M6 : Des contributions plus faibles (environ 2-5 %) par rapport aux chaînes publiques et Canal+.

NRJ12, W9, TMC, TFX : Des chaînes secondaires avec des pourcentages beaucoup plus faibles, souvent inférieurs à 5 %.

Sous-totaux: Les sous-totaux regroupent plusieurs chaînes, ce qui explique leur dominance dans les pourcentages totaux. Points particuliers Certaines chaînes (comme Gulli, France 5) montrent des valeurs isolées, représentant des années ou des moments où elles ont diffusé beaucoup plus ou beaucoup moins de films que leur tendance habituelle.

Les chaînes comme Canal+, France 3, et les sous-totaux montrent une forte variabilité dans leurs pourcentages de diffusion, probablement en fonction de leurs orientations programmatiques annuelles. En résumé Les chaînes publiques et payantes (représentées dans les sous-totaux) sont des acteurs majeurs dans la diffusion de films. Certaines chaînes comme Canal+ et France 3 se démarquent individuellement avec des contributions importantes.

**Comparaison des films français et américains (Bar Chart)** :

Un graphique en barres compare le nombre de films français et américains diffusés par chaîne. On peut remarquer facilement que les chaînes françaises privilégient les films français que les films américains.

**2.3. Conclusion:**

1. La répartition des films varie significativement entre les chaînes publiques, privées, et payantes.
2. Les chaînes françaises privilégient les films français que les films américains.
3. Pour aller plus loin, il serait intéressant d’analyser les genres de films diffusés et leurs audiences respectives.

## Troisième Tableau de bord et Machine learning: Politique

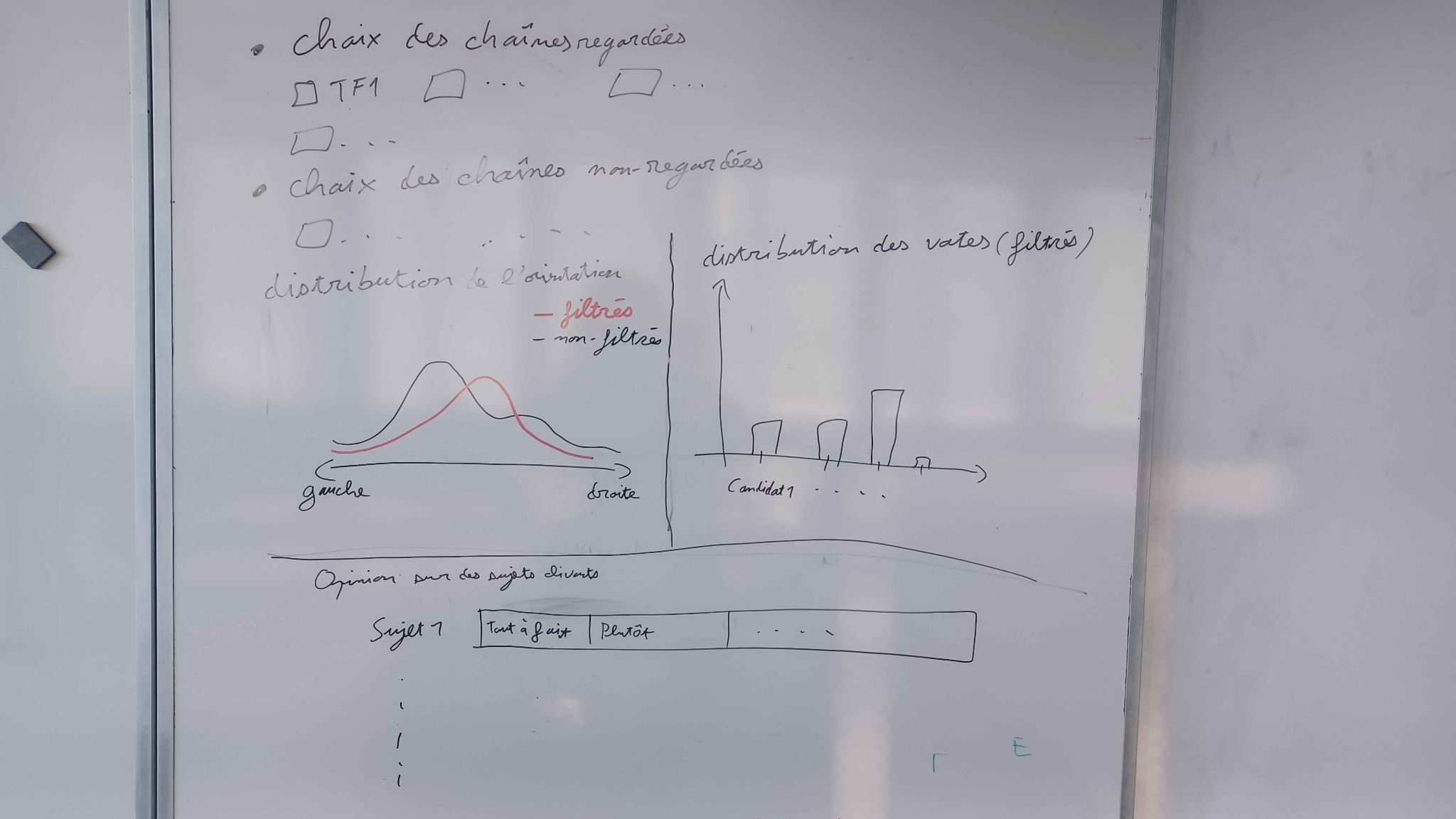
Le dataset utilisé pour ce tableau de bord est celui qui contient les réponses au questionnaire “Enquête sur les français et l’information”. Il est très riche en données mais nous avons dû faire des choix pour nos analyses. Ce tableau de bord se concentre sur l’impact des médias sur les opinions politiques des français.

Les données que nous avons choisies d’exploiter sont les suivantes:

* les chaînes TV regardées par le participant parmi plusieurs propositions;
* une note de 0 (la gauche) à 10 (la droite) indiquant le bord politique déclaré par le participant;
* le vote au premier tour des présidentielles 2022 du participant;
* l’opinion du participant sur 8 sujets clivants parmi “Tout à fait”, “Plutôt”, “Je ne sais pas”, “Plutôt pas”, “Pas du tout”.

A la base les visuels présentés ci-dessous faisaient partie de la phase exploratoire de la partie machine learning (prédiction du bord politique ou du vote à partir des chaînes regardées), mais nous avons finalement décidé d’en faire un tableau de bord.

Le but est de mettre en avant d’éventuelles différences d'opinions politiques entre les spectateurs de différentes chaînes. Le plan du dashboard est présenté ci-dessous.



*Figure 14 : Plan du tableau de bord sur les opinions politiques des spectateur.ices*

Les filtres permettent de choisir quelle chaîne le participant regarde ou ne regarde pas (par exemple on peut choisir un spectateur qui regarde Arte mais pas TF1). Nous avons également ajouté un filtre qui permet de choisir, pour les analyses comparatives, si les données choisies seront comparées à toute la population ou à leurs complémentaires (par exemple si l’on choisit les personnes qui regardent TF1, on peut les comparer à toute la population ou bien à ceux qui ne regardent pas TF1).

Le premier graphe montre la distribution de l’orientation politique (gauche/droite) dans l’échantillon filtré, et la compare à l’échantillon de comparaison choisi. En réalisant cette figure, nous avons ajouté une ligne verticale au niveau des moyennes de chaque échantillon et nous avons quantifié la différence entre les deux distributions à l’aide de la distance de jensen shanon.

Le second montre simplement le nombre de votes pour chaque candidat aux présidentielles 2022 parmi l’échantillon filtré, avec à côté les résultats des vraies élections.

Le troisième détaille l’avis des individus filtrés sur les 8 sujets clivants. Lors de la réalisation du dashboard, nous avons également voulu ajouter une comparaison de ces avis avec l’échantillon de comparaison choisi. Nous avons donc affiché sur une quatrième figure la différence de pourcentage d'individus étant d’accord avec l’affirmation (réponse “Tout à fait” ou “Plutôt”) dans les deux groupes, avec un test d’égalité de proportions permettant de griser l’affichage lorsque les avis ne sont pas significativement différents entre les deux populations.

Le rendu final du dashboard est le suivant:



*Figure 15 : Rendu final du tableau de bord sur les opinions politiques des spectateur.ices*

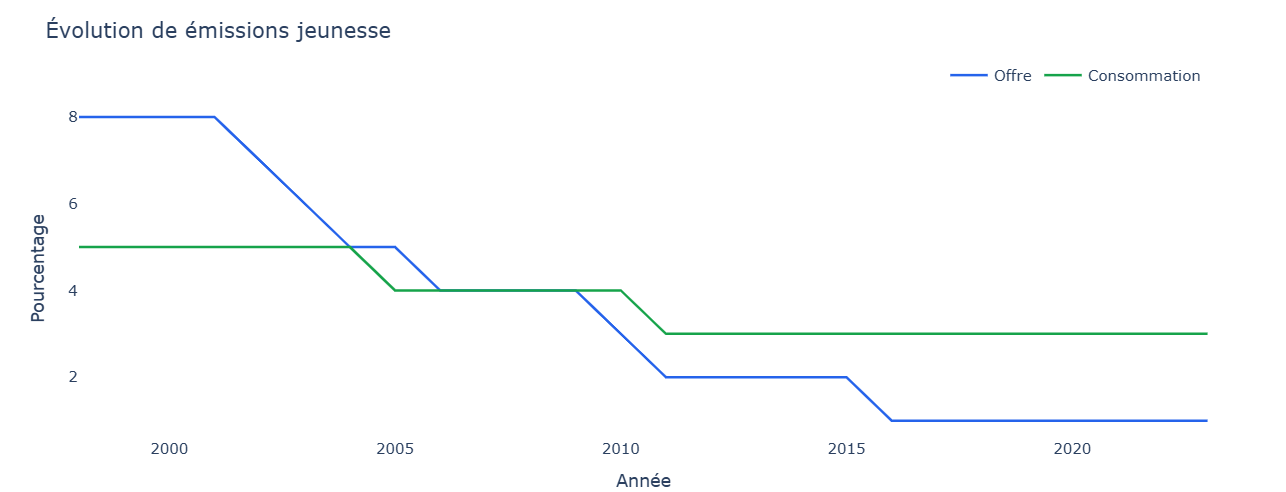
## Quatrième tableau de bord: La consommation en fonction de l’âge

Avant toute chose, nous souhaitons vous faire part de comment nous avons été amenés à cette thématique en particulier.

En parallèle de la conception du dashboard sur les thématiques, nous avons également conçu, à l’aide du jeu de donnée Audience de la télévision.xlxs onglet “O&C”, un dashboard qui évaluait l’offre et la consommation de chaque structure de programme passant à la télévision au fil des années. Les données correspondent au pourcentage d'offres et de consommation pour chaque type de structure en fonction de l’année.

En étudiant les différentes courbes, nous nous sommes rendus compte qu'au fil des années, les consommations et les offres restaient stables en termes de pourcentage. Il est certes possible que les gens regardent moins la télé, mais ce que les gens consomment sur la TV n’évolue pas forcément au fil des années.

Cependant, il existe une seule exception: les émissions de jeunesse. En effet, l’offre et la consommation de ces dernières ont connu une baisse significative au cours des dernières années. Cela a éveillé notre curiosité, nous indiquant un potentiel délaissement des jeunes vis-à-vis de la télévision.



*Figure 16 : Évolution de l’offre et de la consommation des émissions jeunesse*

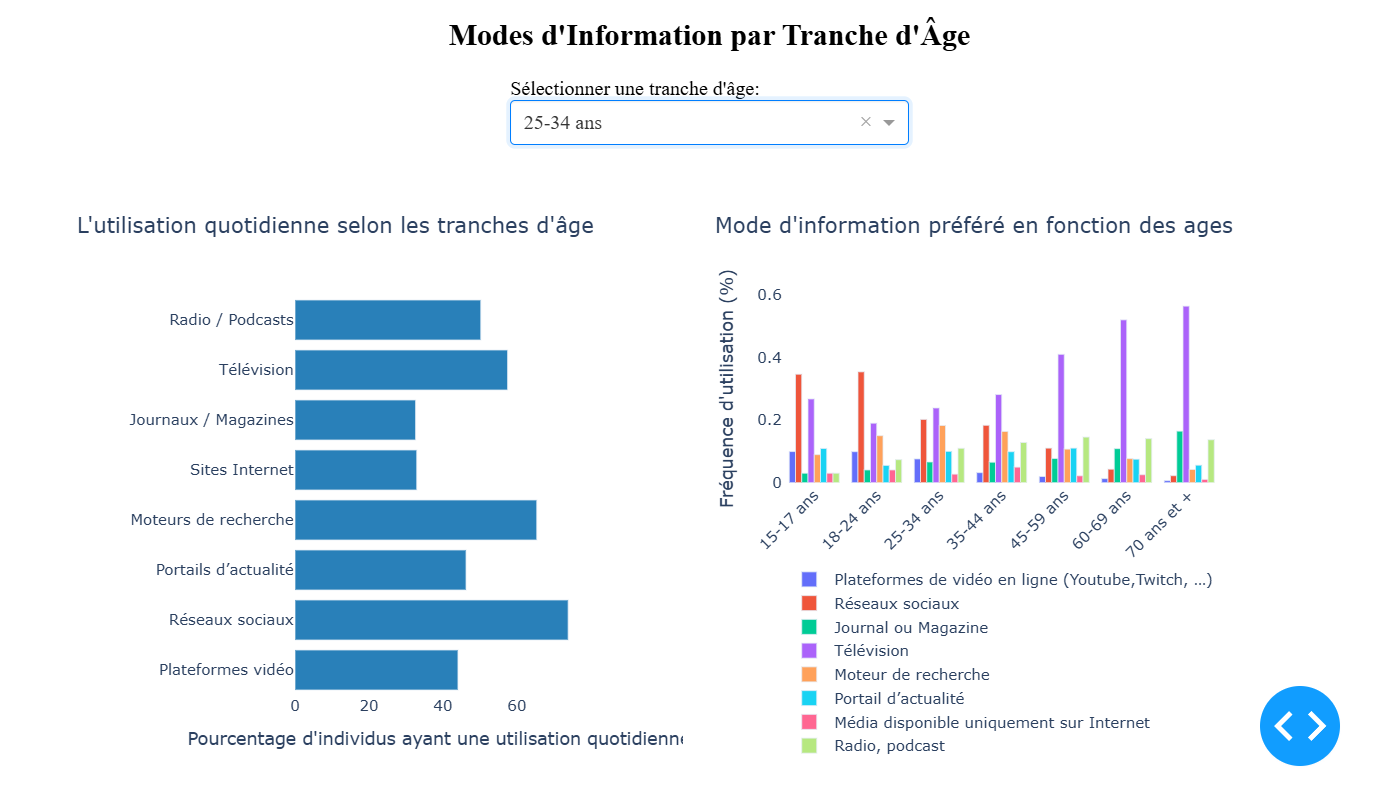
Nous nous sommes donc par la suite penchés sur ces nouvelles problématiques : Quelle est la consommation des jeunes en terme d’audiovisuel? Est-ce que les nouveaux médias ont pris le dessus concernant les jeunes générations? Cela affecte-t-il leur concentration?

Dans un premier temps, nous étions tentés de piocher des jeux de données sur statista.com, car ceux-ci nous auraient permis de réaliser des études temporelles pour voir l’évolution du potentiel désintérêt des jeunes pour la télévision, et le mettre en contradiction avec l’augmentation probable de leur consommation de nouveaux médias. Nous avons pu obtenir ces datasets, bien que payants, mais dans un souci de légalité et de plein respect des lois, nous avons préféré ne pas exploiter ces données.

Cela veut dire que pour la suite, nous nous baserons uniquement sur les réponses du questionnaire de l’Arcom de décembre 2023. Nous n’aurons pas d’évolution temporelle, mais étant donné le caractère récent de cette enquête, nous aurons quand même de bons éléments pour répondre à nos problématiques.

En nous basant sur les 8 colonnes NEWS1\_R qui représentent les fréquence d'utilisation pour chaque modes d'information, nous avons étudié le pourcentage de gens qui utilisent au moins une fois par jour tel ou tel support, et ce pour chaque tranche d’âge. Pour ce faire, nous avons simplement additionner les gens ayant répondu “Au moins une fois par jour” avec ceux ayant répondu “Plusieurs fois par jour” Ce graphique nous permettra potentiellement de voir s' il y a une certaine diversité d’utilisation de mode d’information selon les tranches d’âge, ou au contraire certaines disparités.

Sur la même page du dashboard à côté de ce graphique, en nous basant sur la colonne NEWS1BIS\_R, nous avons fait un graphique en histogramme qui montre le mode d’information préféré pour chaque tranche d’âge, sous forme de fréquence pour chaque groupe d’âge.



*Figure 17 : Visuel de l’onglet Modes d’information*

Cette courbe nous a directement interpellé, en effet les visuels sont très parlant, les jeunes n’ont pas du tout les mêmes sources d’information préférées que les personnes plus âgées.

Nous avons donc directement voulu faire du machine learning sur ces données la. Nous avons utilisé une régression logistique multinomiale, qui est un modèle statistique simple et adapté à ce type de problème de classification avec plusieurs catégories.

L’idée de notre code était de linéariser nos tranches d'âge de sorte à avoir des valeurs continues, et pouvoir ainsi, pour chaque âge, tenter de prédire les probabilité que tel ou tel support soit le moyen d’information préféré de la personne en question. Nous observerons les résultats par la suite.

L'intérêt des jeunes pour les nouveaux types de réseau nous a questionnés sur la possible corrélation avec la baisse de concentration qu’on semble observer sur les jeunes générations. Et par chance, beaucoup de personnes ont répondu sur leur préférence vis à vis du de la longueur des vidéos qu’ils préfèrent consommer ( colonne FORMAT3\_LR2\_FORMAT3\_3\_R ). Nous avons donc procédé, à l’aide d’histogrammes, à une étude par âge, en espérant une préférence des jeunes sur les formats plus courts par rapport aux personnes plus âgées.

Si tel est le cas (spoiler : oui), il serait intéressant de voir le lien avec des réseaux sociaux tels que Tiktok ou Instagram où le format court est très prisé. Les interrogés ont pu répondre, s’ils le souhaitaient, s’ils utilisaient ou non différents réseaux sociaux, et notamment ces deux-la. C'est donc avec ces données-là que nous avons conçu un diagramme en histogramme où pour chaque format préféré pour chaque tranche d’âge, nous voyons la part d’utilisateurs et de non-utilisateurs de Tiktok et Instagram. Les résultats attendus seraient donc une forte utilisation de ces deux réseaux pour les gens préférant les contenus courts.

# Résultats

Cette section revient sur les résultats clés du projet, comprenant la liaison ERP BD TOPO, la classification du risque des bâtiments accompagnée de la cartographie du risque, ainsi que la proposition d’une nouvelle métrique.

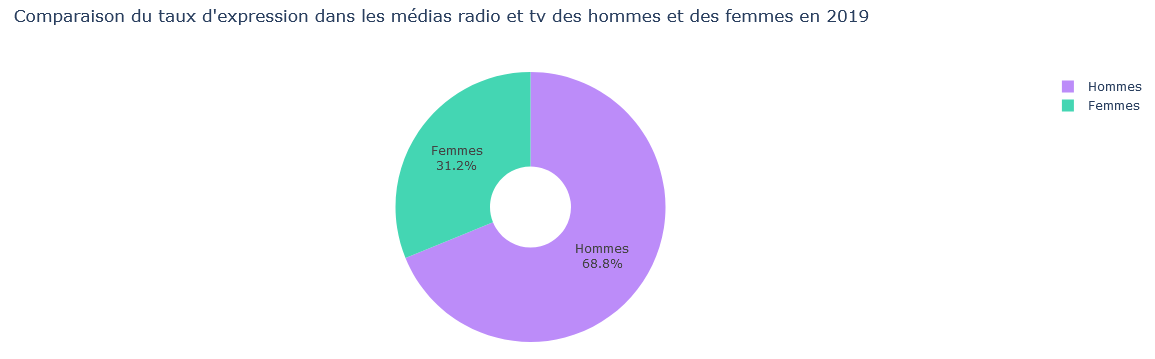
## Les thèmes abordés lors des JT

Concernant les cinq chaînes présentes dans les données fournies, on remarque une forte proximité entre quatre d’entre elles: TF1, France 2, France 3 et M6, tandis que Arte semble traiter des thèmes assez différents, plutôt axés sur l’International et les thèmes “Culture/Loisirs”.  
Au global, nous avons remarqué une hausse vertigineuse du thème de la santé en 2020, qui s’explique logiquement par l’arrivée mondiale de la pandémie du COVID, cependant de nombreux thèmes ont connu des pics de popularité, comme par exemple le thème “Catastrophe” en 2010, année très marquée par les catastrophes naturelles (300 000 morts, bilan le plus meurtrier de l'histoire depuis 1983).

Pour certains autres thèmes on observe une saisonnalité des résultats: tous les cinq ans en politique, correspondant aux élections présidentielles et tous les deux ans en sport, correspondant aux coupes d’Europe et du monde de football.

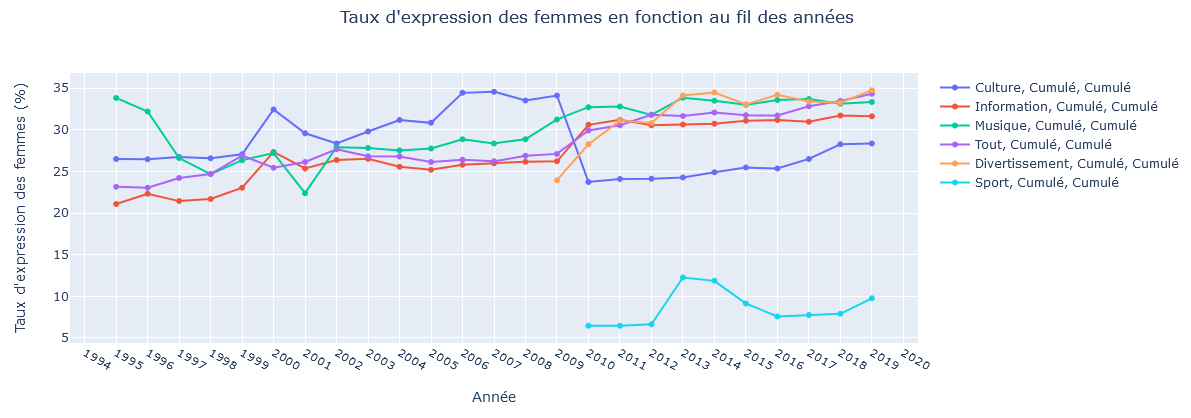
## Le temps de parole des femmes à la télévision et à la radio

Tout d’abord on voit qu’en moyenne sur toutes les années (1995 à 2019) et toutes les chaînes étudiées (radio et tv) les femmes ont parlé 29.2% du temps. Les hommes ont donc parlé 70.8% du temps. En 2019, la dernière année étudiée, la répartition est de 31.2% pour les femmes et 68.8% pour les hommes. Il y a donc de grandes inégalités sur le temps de parole au global. Nous allons donc regarder cela plus en détail.



*Figure 18: Taux d’expression des femmes en 2019*

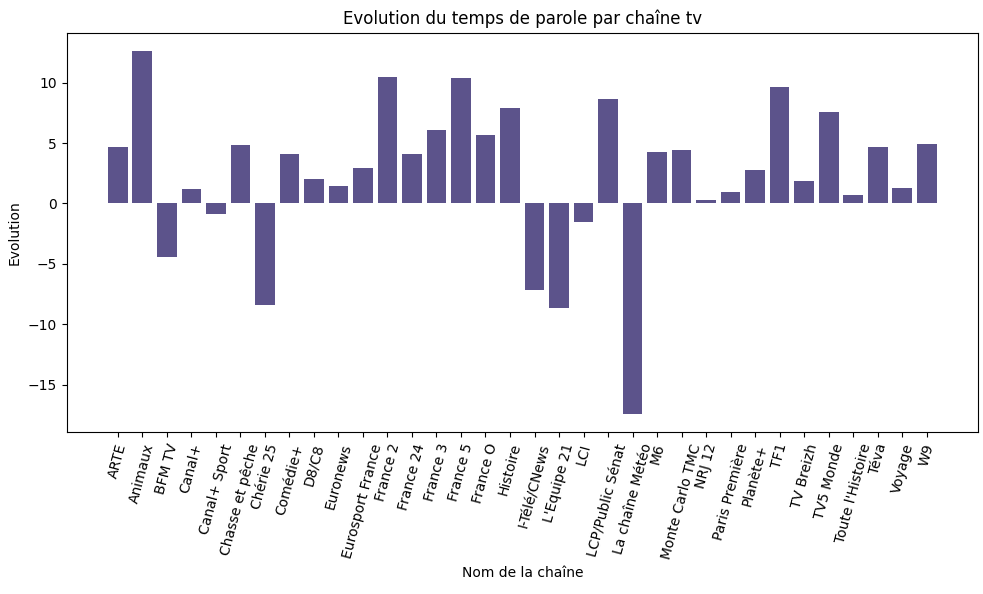
Si l’on sépare les chaînes en 6 thèmes on remarque que c’est dans le thème du sport que les femmes sont le moins bien représentées avec à peine 10% de taux de parole. Les autres thèmes varient entre 28% et 35% en 2019.



*Figure 19: Évolution du taux d’expression des femmes*

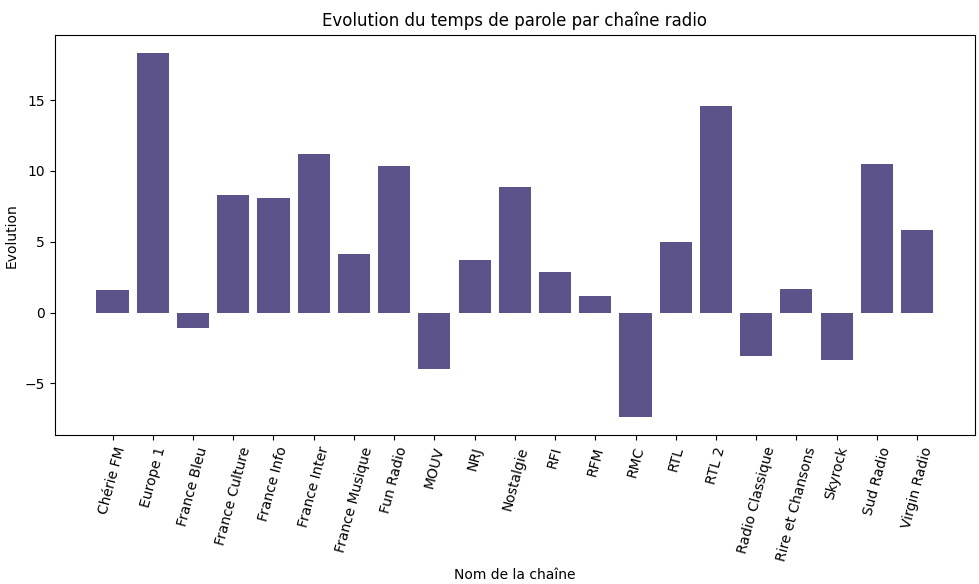
Pour analyser plus en détail ces évolutions, nous avons représenté la différence entre le taux d’expression des femmes en 2019 et en 2010 (ou en 1995 pour les stations de radio) pour chaque chaîne et station. Cette visualisation permet de dresser un bilan de l’évolution pour chacune d’elles.

On observe ainsi que la chaîne ayant connu la plus forte progression est Animaux (+13 %), tandis que La Chaîne Météo affiche la pire performance (-17%). Globalement, la plupart des chaînes ont une évolution inférieure à 10%.



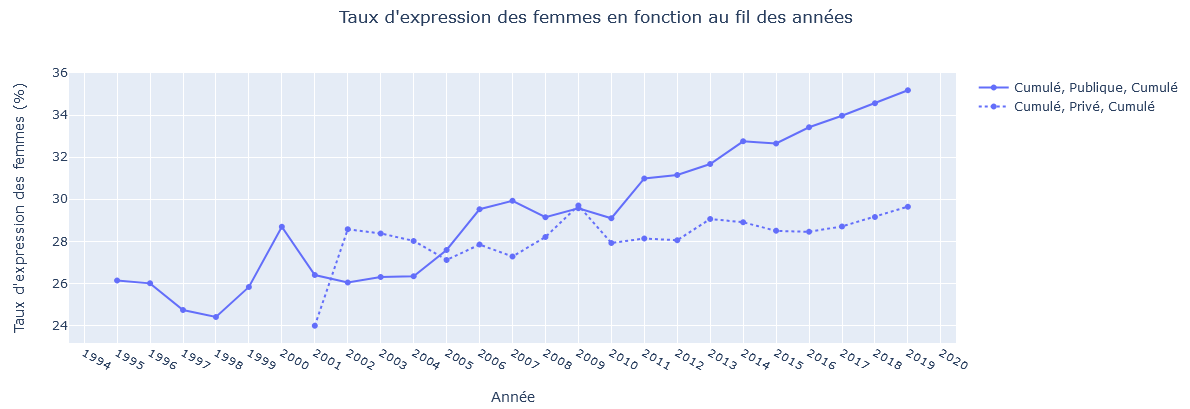
*Figure 20: Évolution du taux d’expression des femmes par chaîne TV de 2010 à 2019*

Pour ce qui est des stations radio, Europe 1 est la station avec la plus grande évolution (+19%).



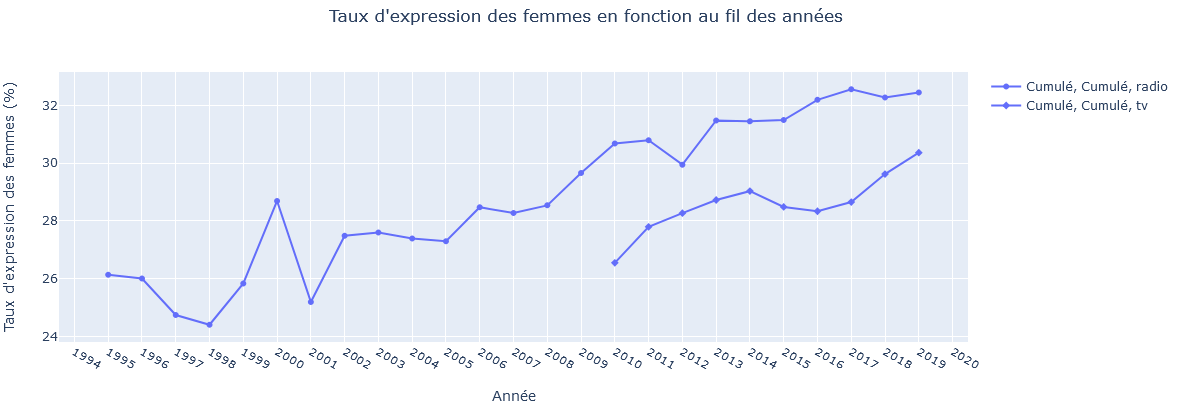
*Figure 21: Évolution du taux d’expression des femmes par radios de 2010 à 2019*

On peut aussi différencier les chaînes selon si elles sont publiques ou privées. On voit que globalement les chaînes privées ont toujours été en dessous des chaînes publiques et en 2019 il y a un écart de presque 6% avec 29.6% de taux de paroles des femmes sur les chaînes privées et 35% pour les chaînes publiques. Cela peut s’expliquer par le fait que dans notre échantillon de chaînes, les chaînes de sport sont toutes privées et elles ont un taux très faible.



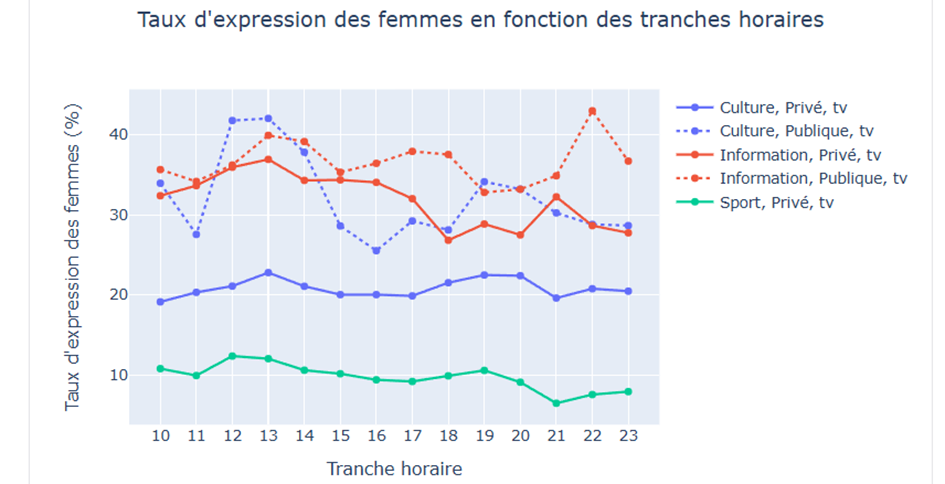
*Figure 22: Évolution du taux d’expression dans le privé et le public*

Pour finir, il n’y a pas tellement de différences entre la télévision et la radio, la radio a 3 ou 4% de plus que la tv selon les années. Cependant si on regarde l'évolution, il semblerait que la télévision augmente petit à petit depuis 2016 (2% de plus en 2019) alors que la radio a bien augmenté depuis ses débuts jusqu’à 2013 mais depuis elle stagne (seulement 1% de plus en 6 ans).



*Figure 23: Évolution du taux d’expression des femmes à la,TV et la radio*

A l’aide du graphique illustrant le taux de parole en fonction des tranches horaires, nous avons essayé de voir si pour les chaînes ayant un thème similaire, on pouvait observer des fluctuations notables dans le pourcentage.



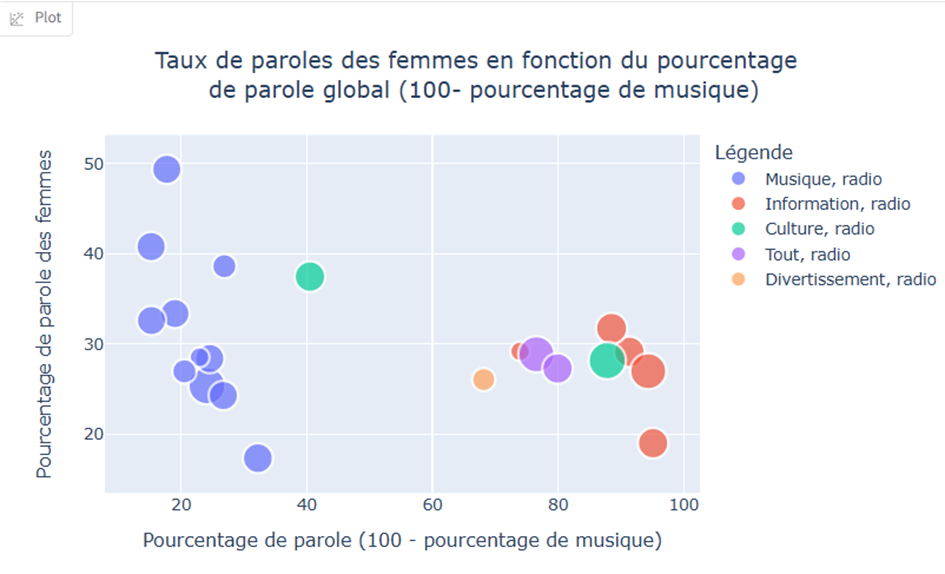
*Figure 24: Taux d’expression des femmes par heure*

En faisant un focus sur les thèmes Culture, Information et Sport ainsi que sur les chaînes TV, on remarque à nouveau que le thème Sport est le moins représenté par des femmes. En effet, ici pour le thème Sport, le taux de représentation stagne autour des 10 % pour atteindre un minimum de 6.5% à 21h (heure des plus fortes audiences).

Cette même tendance se retrouve aussi sur les chaînes culturelles privées où le taux stagne autour des 20% avec une décroissance de 3.5% sur les heures de forte audience.

Pour les chaînes TV culturelles publiques, la décroissance se constate aussi à partir de 20h (5% de diminution) même si en moyenne sur l’ensemble de la journée le taux d’expression reste beaucoup plus important que sur les chaînes privées (écart d’environ 15% avec les chaînes privées).

A présent, nous allons évoquer ce que nous avons pu retirer du graphique représentant le taux de parole des femmes en fonction du pourcentage de parole global du média. En ne filtrant uniquement sur les médias radio, on remarque plusieurs choses. Tout d’abord, la distinction entre les chaînes musicales et les autres radios est nettement visible.

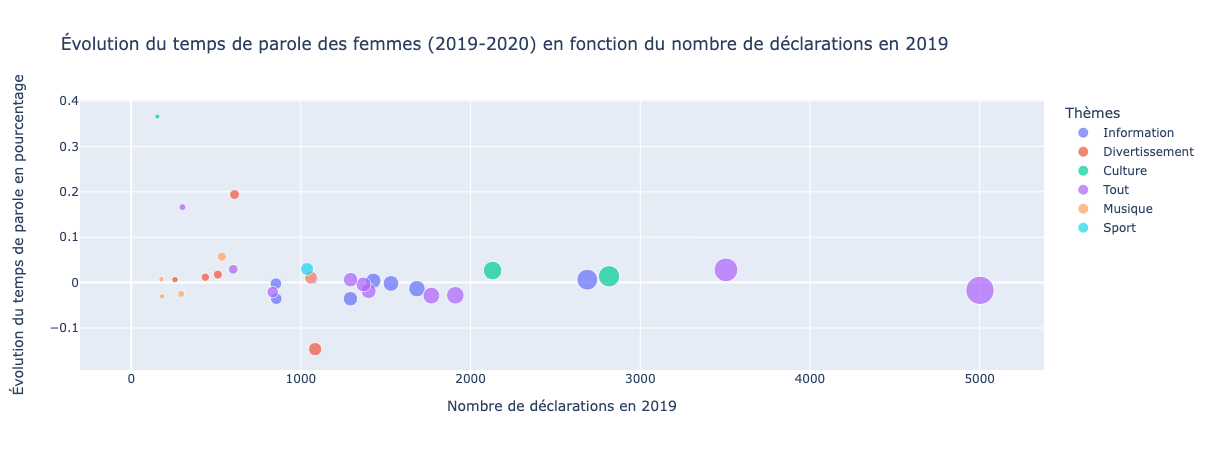


*Figure 25: Temps de parole des femmes selon le temps total de parole*

De plus là où pour les chaînes musicales on en compte trois donc le pourcentage de parole des femmes est au-dessus de 38% (radio classique, RTL2 et Chérie FM), pour les chaînes non musicales, l’ensemble des chaînes est en dessous des 32% de parole mise à part la chaîne culturelle France Bleu qui atteint les 37.5 % de pourcentage de parole.

Par ailleurs, là où la distribution pour les radios musicales est beaucoup plus large (de 17.3 % à 49.3 %), celle des radios non musicales est beaucoup plus resserrée (entre 19% et 37.5 % dont 80% de ces radios entre 26 et 32%.

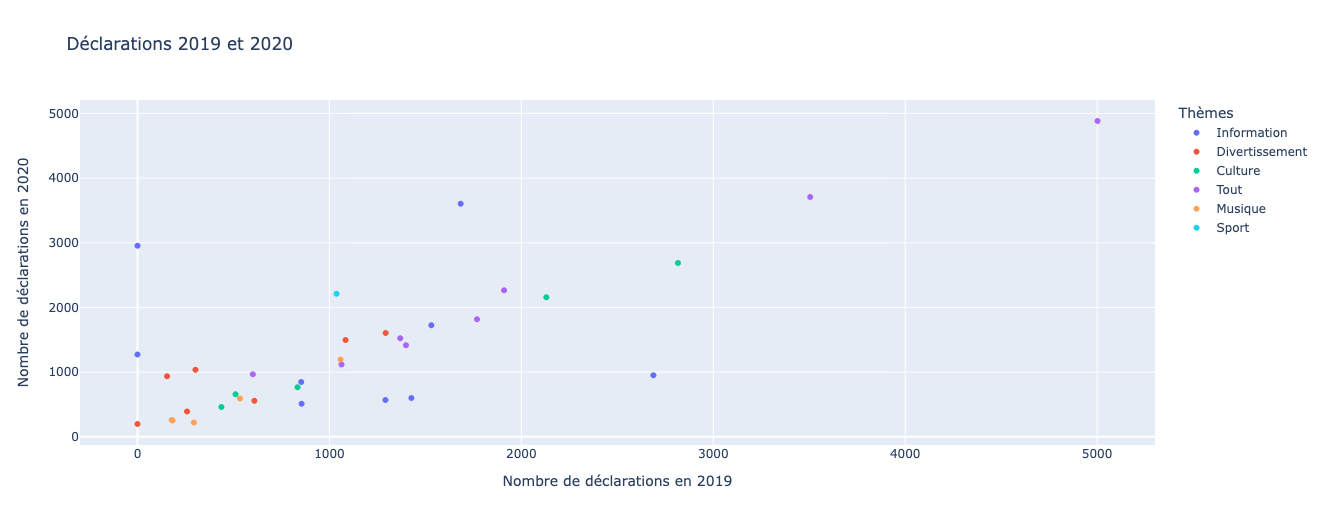
Enfin, nous avons effectué des graphiques sur les données des chaînes ayant les programmes qui ont fait l’objet d’une déclaration. Le premier visualise l’évolution du pourcentage de temps de parole des femmes en fonction du nombre de déclarations en 2019 pour chaque média.



*Figure 26: Impact du nombre de déclarations au CSA sur le temps de parole féminin*

Sur ce graphique, **14** médias présentent une baisse du taux d’expressivité des femmes malgré leur nombre de déclarations. En moyenne, l’évolution du taux d’expressivité entre 2019 et 2020 est d’environ **1%**. Pour la majorité des médias, la différence du taux d’expressivité entre les deux années oscille entre **-5%** et **5%**, bien qu’on note une forte évolution pour France 4, avec une hausse de **36%.** À l’inverse, Chérie 25 présente une baisse importante de **14%**. Les caractéristiques des médias (public ou privé, thèmes abordés) n’apportent pas de conclusions significatives.

Le second explore la corrélation entre le nombre de déclarations en 2019 et en 2020 pour chaque média.



*Figure 27: Nombre de déclarations au CSA en 2019 vs 2020*

Sur le second graphique, une forte corrélation **0,75** est observée entre le nombre de déclarations en 2019 et en 2020. En général, les médias ayant déclaré un grand nombre de programmes en 2019 continuent de le faire en 2020. Une analyse par thème révèle une différence notable entre les déclarations de 2019 et 2020 pour les médias classés dans la catégorie **Information**.

## La différence d’âge entre les acteurs et les actrices des films ayant fait le plus d’audience.

Tout d’abord d’un point de vue global, sur toutes les chaînes de 2002 à 2023 , nous pouvons constater qu’en moyenne les femmes sont significativement plus jeunes que les hommes dans les films faisant le plus d’audience, et ce indifféremment de la note au test de bechdel. Nous voyons aussi que cette moyenne est différente pour les films avec une note de 0(rouge) qui montre une différence d’âge significativement supérieure de 11 ans, c’est à dire que l’actrice principal est en moyenne plus jeune d’au moins 11 ans. L’autre groupe de film avec des notes entre 1 et 3 a une moyenne assez similaire mais supérieure à 4 ans.

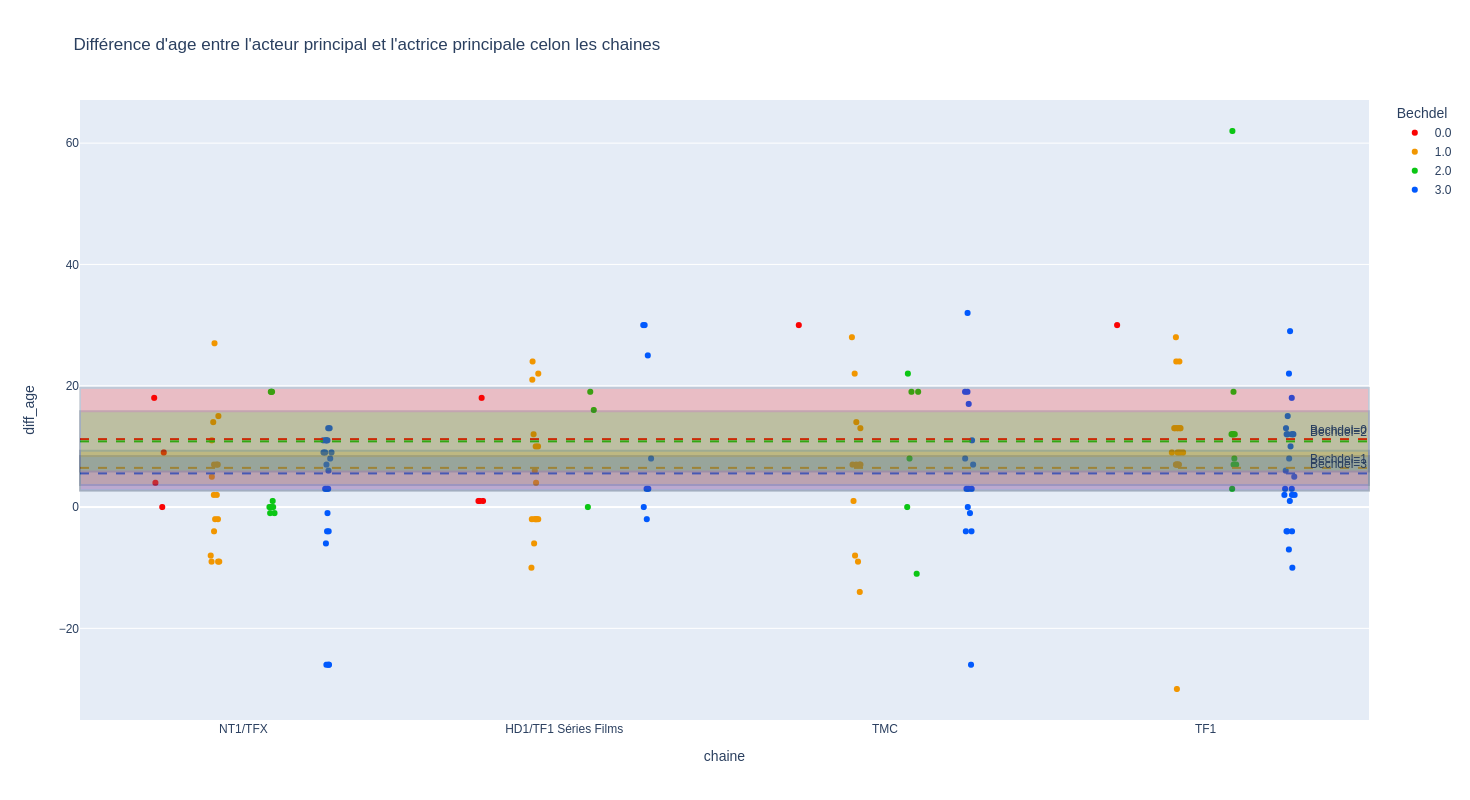
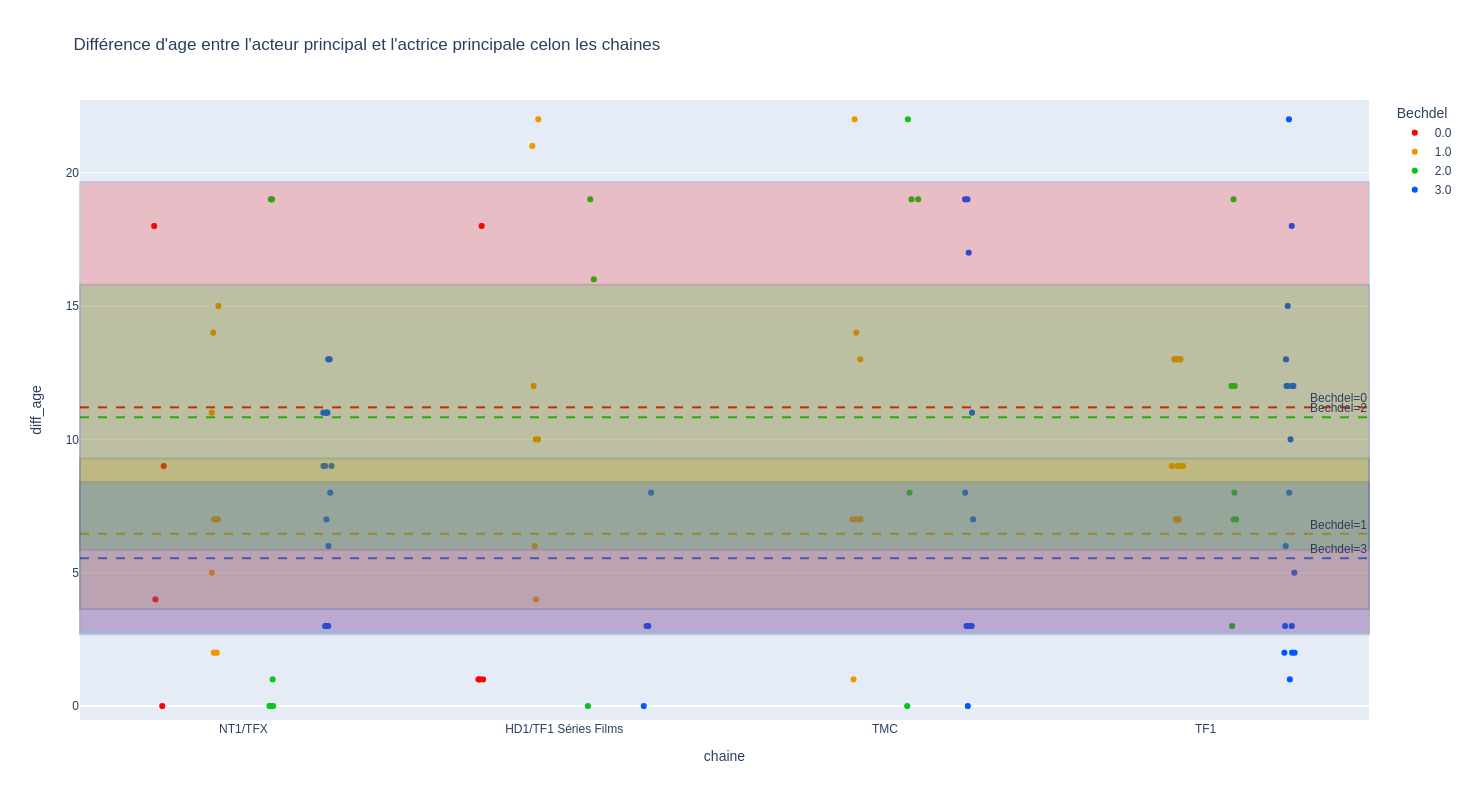
 *Figure 28 : Zoom sur la différence d’âge entre acteurs et actrices par chaînes*

La continuité de l’étude aurait dû être d'observer des différences de répartition des

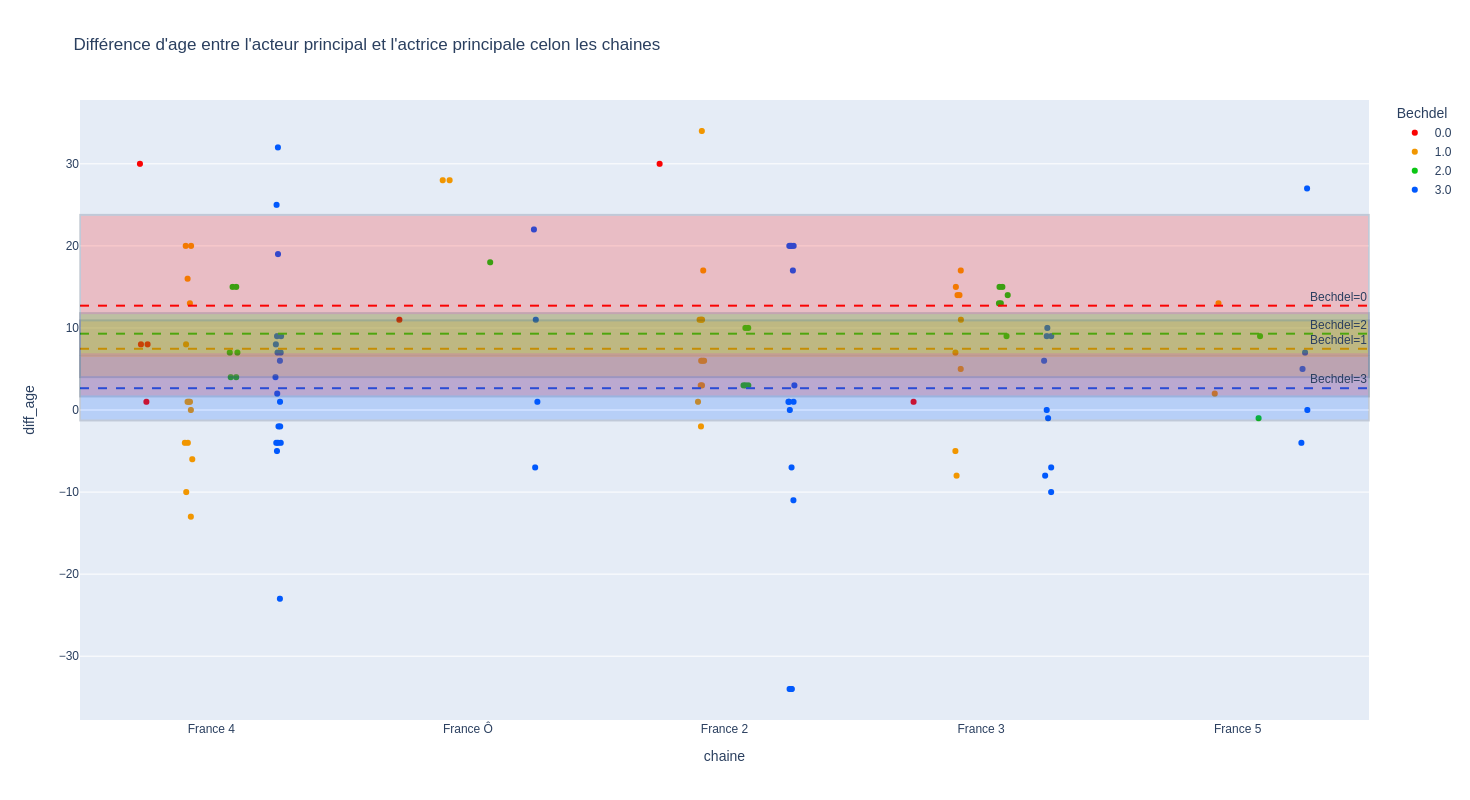
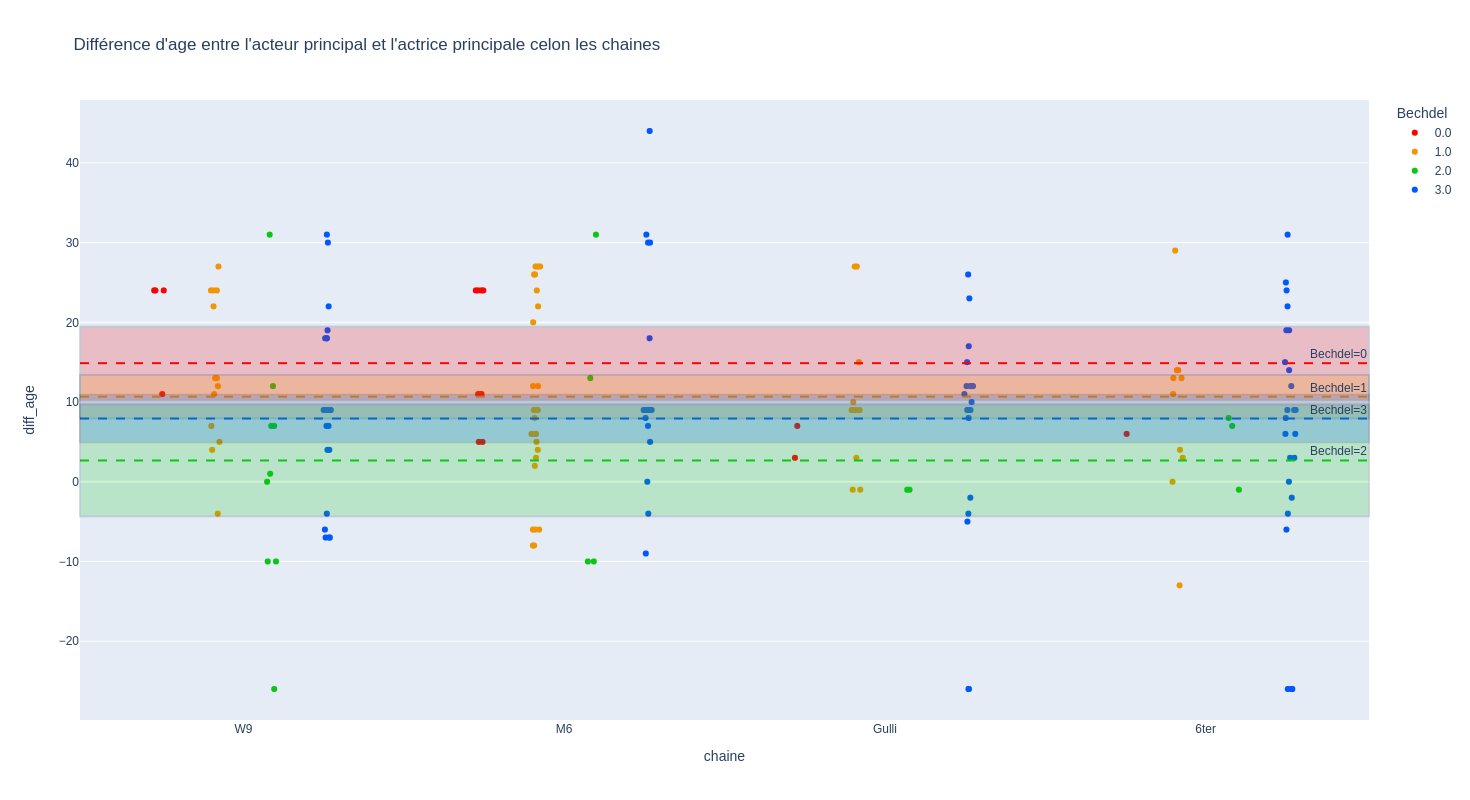
notes et des différences d’âges entre les chaînes mais il y a eu un problème de jointure entre les 2 jeux de données. Le problème était qu’un dataset avais les tire des films en français tandis que l’autre les avais en englais, la solution a été de faire scraping pour pouvoir traduire les titres en englais mais cette étape a pris du temps pour être mise en place et nous a fait perdre la moitié des données. La perte de temps ne nous a pas permis de créer des groupes de chaîne qui aurait la même répartition, et la perte de donnée nous limite et ne permet pas de faire voir de bonne représentation par chaîne ou par plus petite période. Malgré cela nous allons tout de même tenter de voir de particularité qui peuvent se dégager des groupes de média et si elle diffère d’un groupe à l’autre de 2002 à 2023.

Nous allons donc nous intéresser au trois groupe les mieux représenté, le groupe TF1, le groupe M6, et France Télévision.

Pour le groupe TF1, il n’y a pas de différence notable concernant la répartition des notes de bechdel entre les chaînes de ce groupe, on notera qu’il y a assez peu de films avec un bechdel égal à 0 ou 2. Concernant la différence d’âge, on ne peut pas dire que les moyennes soit différente d’une note à l’autre, mais elle sont toute significativement plus grande que 0, dans comme globalement l’actrice est plus jeune que l’acteur.

*Figure 29 : Différence d’âge sur le groupe TF1 (zoom à droite)*

Ensuite via à vis du groupe M6 et de France télévision, nous faisons les mêmes constatations que précédemment, en ajoutant que M6 et W9 sont très proches dans la répartition, ce qui s'explique par la grande similarité des films faisant le plus d’audience, W9 devant surement avoir une programmation très similaire.

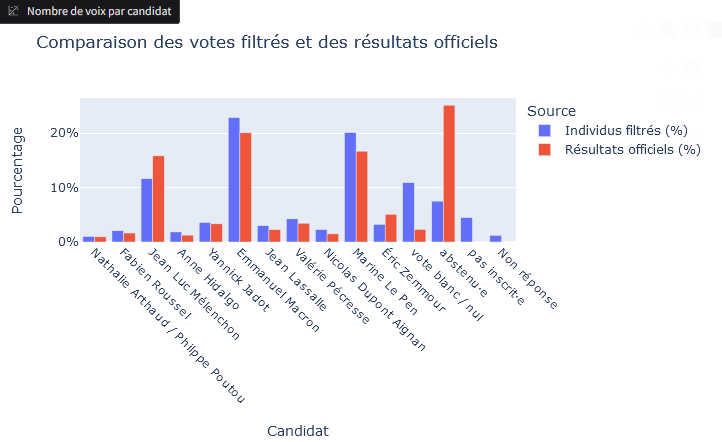
*Figure 30 : Différence d’âge sur le groupe M6 à gauche, sur France Télévision à droite*

Pour conclure nous n’arriverons pas remarquer des différences entre les groupes de chaîne. Donc les chaînes n’ont possiblement pas d’influence ni sur la note au test de bechdel films faisant le plus d’audience ni sur la proportion de note, du moins nous n’arrivons pas à le remarquer.

## L’impact des médias sur l’opinion politique

Comme nous l’avons vu, même si la tendance est à l’amélioration, les femmes sont toujours très mal représentées sur les chaînes de télévision française et à la radio. Nous voulons à présent montrer que les images véhiculées par la télévision sont importantes et ont un impact sur les opinions et décisions politiques de tout un chacun. Le tableau de bord que nous avons développé sur ce sujet va nous aiguiller.

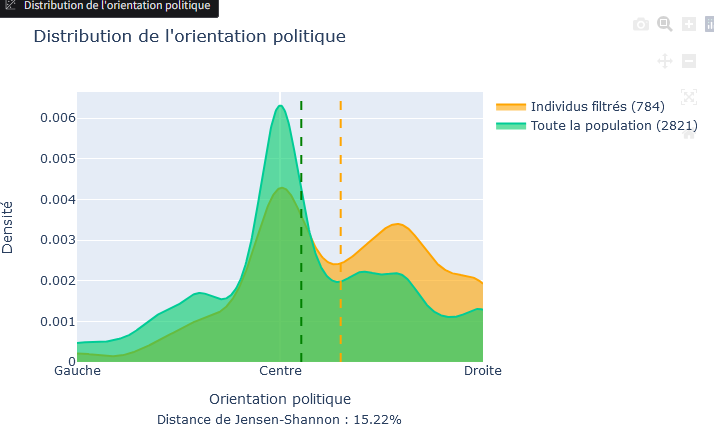
Une première observation à faire est que lorsque l’on regarde l’ensemble de la population, les individus étudiés sont plutôt à droite, autant dans leur dire que dans leur vote. On peut même voir sur la figure 31 que les votes de l'échantillon sont davantage à droite que dans les résultats réels de l’élection.



*Figure 31 : Vote de l’échantillon entier aux présidentielles 2022*

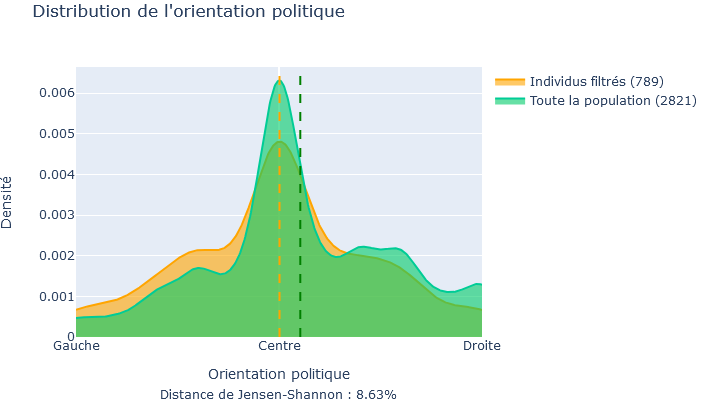
Les tendances générales sont les suivantes: les français pensent en grande majorité qu’il faut aller plus loin dans l’égalité homme femme (89.61 %), que l’on doit changer nos modes de vies pour lutter contre le réchauffement climatique (88.28 %) et que les juifs sont des français comme les autres (83.81 %). Les autres sujets clivent davantage la population, bien que sur ces derniers, ce soit les opinions de droite qui l’emportent la plupart du temps, excepté sur l’affirmation “il est normal que les couples homosexuels puissent avoir des enfants”.

On ne va pas donner le détail de toutes les chaînes ici, mais on peut comparer le public de chaque chaîne à la population. On peut prendre le cas des spectateurs de CNews qui est le plus marquant. Parmi les chaînes citées, c’est celle qui s’éloigne le plus de la moyenne, avec une distance de Jensen-Shannon à 15.22 %. Sans surprise, cet éloignement se fait vers la droite, et le vote pour Marine Le Pen y est bien plus élevé que dans les résultats réels de l’élection, a contrario du vote pour Jean-Luc Mélenchon. Les votes pour Emmanuel Macron sont similaires. Les opinions de ces spectateurs sur tous les sujets sont significativement plus à droite que dans la population totale, à l’exception de la phrase “Les juifs ne sont pas des français comme les autres”.



*Figure 32 : Distribution de l’orientation politique pour les spectateurs de CNews*

Un tel écart du côté de la gauche n’existe pas dans le dataset. La chaîne citée la plus à gauche est Arte avec une distance de Jensen-Shannon à 8.63 %. Cependant il serait plus correct de dire que le public de cette chaîne est moins à droite, car les votes perdus pour Marine Le Pen se reportent plutôt sur Emmanuel Macron que sur des candidats de gauche. Les opinions significativement différentes portent sur l’immigration, l’avortement et la question des juifs, avec un penchant à gauche comme on s’y attend. En revanche, sur les autres sujets, on n’observe pas de changements significatifs.



*Figure 33 : Distribution de l’orientation politique pour les spectateurs d’Arte*

En observant ces données pour toutes les chaînes, on peut les catégoriser selon si les spectateurs s’estiment à, votent à et ont des opinions de droite ou de gauche. Les chaînes dont les spectateurs sont plutôt de droite sont TF1, BFM TV, CNews, C8, RMC Story et LCI. Celles de gauche sont France 2, France 3, France 5, Arte et FranceInfo TV. Les spectateurs des chaînes restantes (TMC, Canal+, M6 et France 24) ont des opinions de droite sur certains sujets et de gauche sur d'autres, et leur distribution est similaire à celle de la population totale.

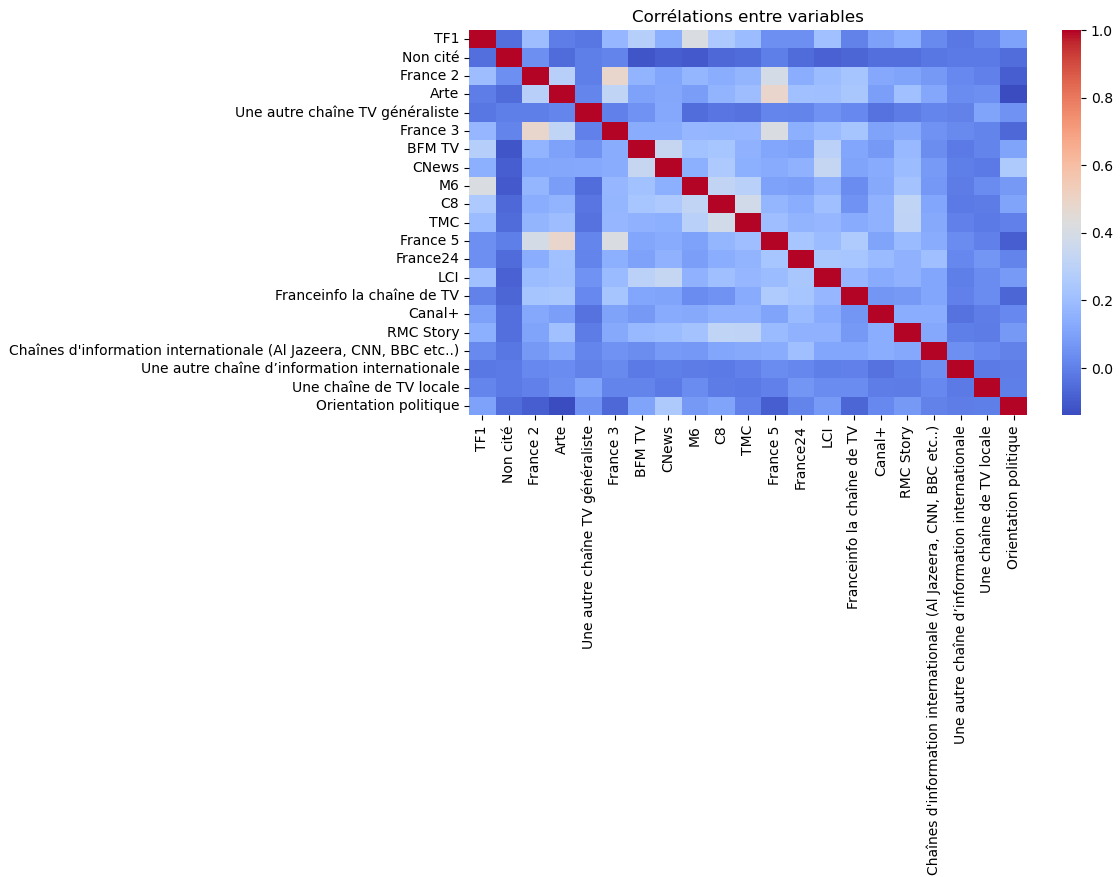
Le dashboard permet aussi d’exclure des chaînes, on peut s’amuser par exemple à regarder les français qui ne regardent que CNews, ce qui, comme on s’y attend, fait pencher drastiquement la balance à droite.

En complément de ces analyses, nous avons voulu voir s’il était possible de prédire l’orientation politique des individus de cette étude à partir des chaînes de TV regardées.Nous avons donc implémenté quelques algorithmes de Machine Learning. Trois méthodes ont été tentées:

La première tentative a été la prédiction du vote aux élections présidentielles 2022 vu comme une classification multiclasse. On a un faible nombre d’individus (2688) et les classes sont fortement déséquilibrées. Après une recherche par grille sur plusieurs algorithmes, le meilleur algorithme qui ressort est xgboost avec un f-score à 0.21 sur l’échantillon de test (580 individus) ce qui est assez mauvais. Ce n’est pas surprenant étant donné la faible population. Nous nous sommes demandés comment améliorer cette prédiction.

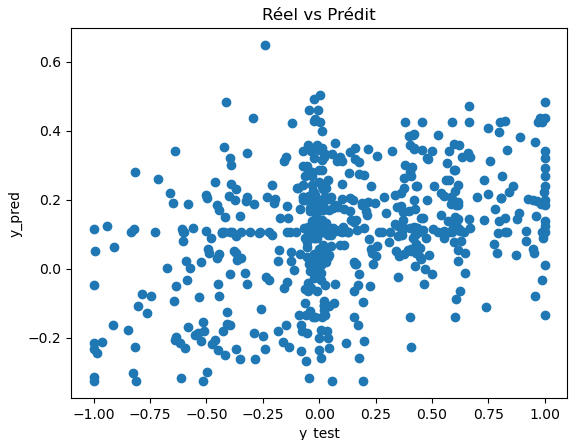
Un des problèmes de la classification multi-classe est l’absence de relation d’ordre entre les classes. En effet, on souhaiterait faire en sorte que prédire “Marine le Pen” au lieu de “Eric Zemmour” soit moins pénalisé que prédire “Jean-Luc Mélenchon”. Nous avons donc pris la décision de changer le problème et de prédire non pas le vote mais l’orientation politique notée entre 0 (gauche) et 10 (droite) des participants. Nous avons ramené ce score entre -1 et 1 et ajouté un léger bruit gaussien autour pour rendre les données continues afin d’appliquer diverses méthodes de régression. On a également pu augmenter nos données à 2899 car plus d’individus ont répondu à cette question qu’à celle de leur vote.

La matrice de corrélation nous montre qu’il n’y a pas ou peu de corrélations linéaires entre le fait de regarder les différentes chaînes et l’orientation politique. Nous avons tout de même tenté d’appliquer plusieurs algorithmes.



*Figure 34: Matrice de corrélation entre les chaînes et ‘orientation politique*

A nouveau, les performances obtenues sont assez mauvaises. Le meilleur algorithme est une régression au K plus proches voisins qui fait une MSE à 0.17, soit 0.41 en RMSE. Pour des valeurs entre -1 et 1 c’est une erreur assez énorme. La figure ci-dessous affiche les prédictions en fonction des valeurs réelles. Pour un prédicteur idéal, les points devraient être alignés sur la diagonale. On en est assez loin, le modèle prédit majoritairement des valeurs entre -0.3 et 0.5. C’est dû au fait que beaucoup de gens dans les données d’entraînement se déclarent au centre.



*Figure 35: Prédiction vs Réalité pour un régresseur KNN*

Nous avons alors décidé de simplifier encore le problème en un problème de classification binaire. Si l’on a du mal à prédire l’orientation politique, on peut éventuellement prédire la tendance gauche/droite d’un individu. On va classer à gauche les personnes ayant répondu 4 ou moins au questionnaire, à droite les personnes ayant répondu 6 ou plus. Pour les personnes ayant répondu 5, deux approches ont été tentées. Dans un premier temps, nous avons tenté de les classer au hasard et dans un second temps, nous avons utilisé le vote aux élections présidentielles pour la classer à gauche ou à droite. Dans le cas où la personne n’a pas répondu à cette question, s’est abstenue ou a voté blanc, elle a été classée aléatoirement. Nous avons décidé de nous baser sur le f-score pour tuner les paramètres des modèles testés car, bien que les classes soient déséquilibrées, il n’y a pas de raison particulière de favoriser l’une plutôt que l’autre.

Là encore les résultats restent assez mitigés bien que meilleurs. On obtient les scores et matrices de confusion suivantes pour chacune des méthodes:

| classement aléatoire des 5 | classement par vote en 2022 des 5 |
| --- | --- |
| **effectif classe 0:** 1678  **effectif classe 1:** 1221  **meilleur modèle:** XGBoost  **f-score:** 0.75 | **effectif classe 0:** 1061  **effectif classe 1:** 1838  **meilleur modèle:** SVM à noyau gaussien  **f-score:** 0.78 |

*Figure 36: Résultats de la classification binaire*

Au vu des effectifs, on peut déjà noter qu’utiliser le vote de la personne qui se déclare au centre l’envoie plutôt dans le groupe de droite, probablement à cause du vote Macron que nous avons placé à droite. On obtient un meilleur résultat en utilisant le vote. Cependant le nombre d'erreurs sur la classe 0 reste conséquent.

Au vu de tout ceci, il n’est pas si simple de prédire l’orientation politique des gens uniquement en utilisant les chaînes regardées. On observe tout de même quelques tendances que l’on peut expliquer à l’aide du dashboard précédent, car comme on l’a vu, le public de certaines chaînes comme CNews fait pencher la distribution à droite. Bien sûr on imagine que d’autres facteurs entrent en jeu dans l’orientation politique des individus, et on pourrait enrichir le modèle avec des réponses à d’autres questions comme les journaux lus, les radios écoutées ou les chaînes youtubes regardées. Celà dit, plus on inclut de variables explicatives, plus un modèle de machine learning a besoin de données, et nous n’en disposons que d’un faible nombre. Nous avons donc choisi d’en rester là.

## La consommation d’audiovisuel chez les jeunes

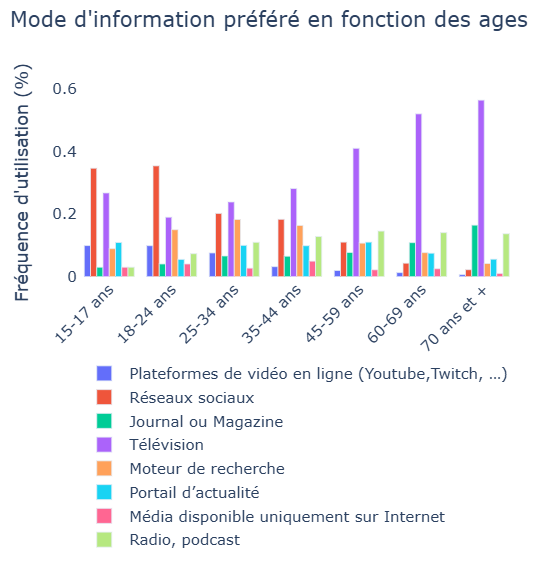
Comme nous l’avons vu précédemment dans la présentation de ce dashboard, il semblerait que tous les types de structures ont des consommations et des offres stables en termes de pourcentage au fil des années. Il est certes possible que les gens regardent moins la télé, mais ce que les gens consomment sur la TV n’évolue pas forcément au fil des années. Il y a juste les émissions jeunesse qui passent dont les pourcentages ont diminué au fil du temps.

Concernant les fréquences quotidiennes des modes d’informations, il semble y avoir une forte diversité chez les jeunes générations dans les médias utilisés, qu’ils soient anciens comme la télévision ou nouveau comme les réseaux sociaux. En revanche, l’usage quotidien de ces supports d'information est beaucoup moins diversifié chez les personnes âgées, préférant beaucoup plus la télévision que les plateformes vidéo par exemple.

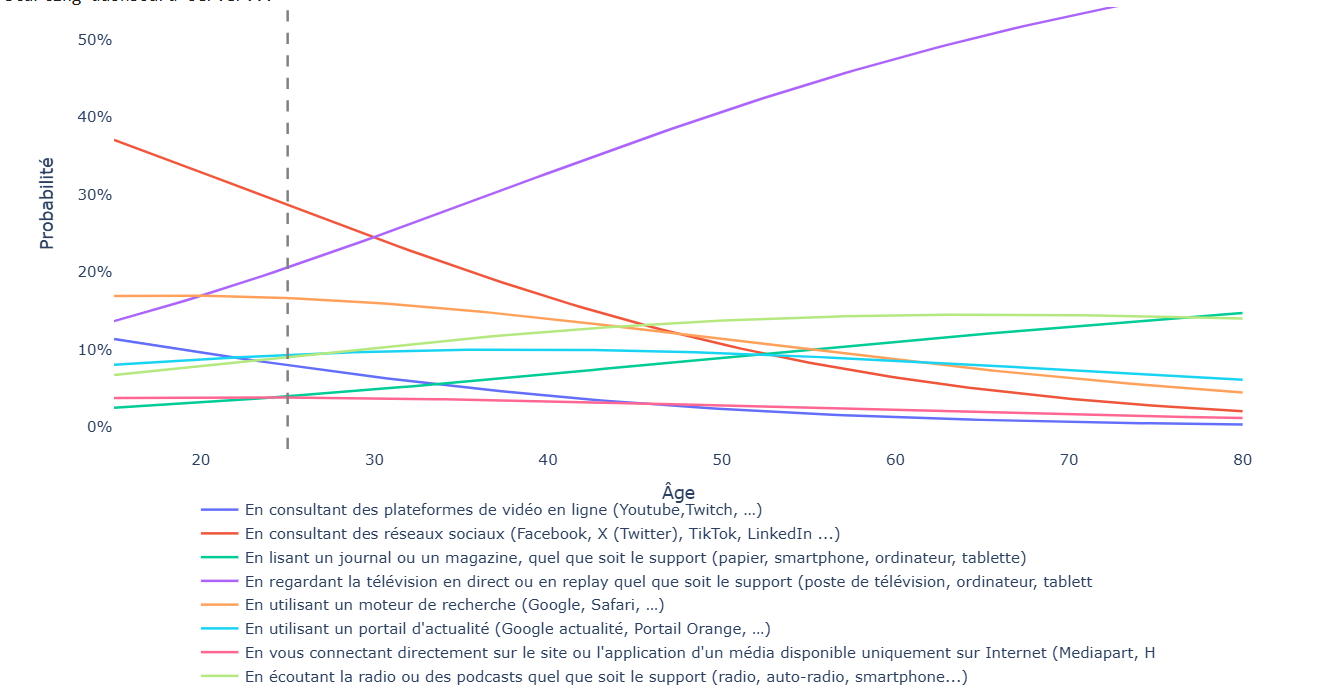
## 

*Figure 37: Fréquence des 18-24 ans (à gauche) et des 60-69 (à droite) qui utilisent quotidiennement tel ou tel mode d’information*

Concernant les préférences de chaque génération, les différences sont sans appel. Les jeunes ont une préférence pour les réseaux sociaux, tandis que les plus âgés préfèrent de façon écrasante la télévision. Et cette tendance se retranscrit très bien sur notre graphe de prédiction. A partir de 30 ans, le mode d’information préféré de l’individu en question est plus probablement la télévision, tandis qu' avant, il s’agit des réseaux sociaux. La domination de la télévision sur les autres médias est de plus en plus importante.

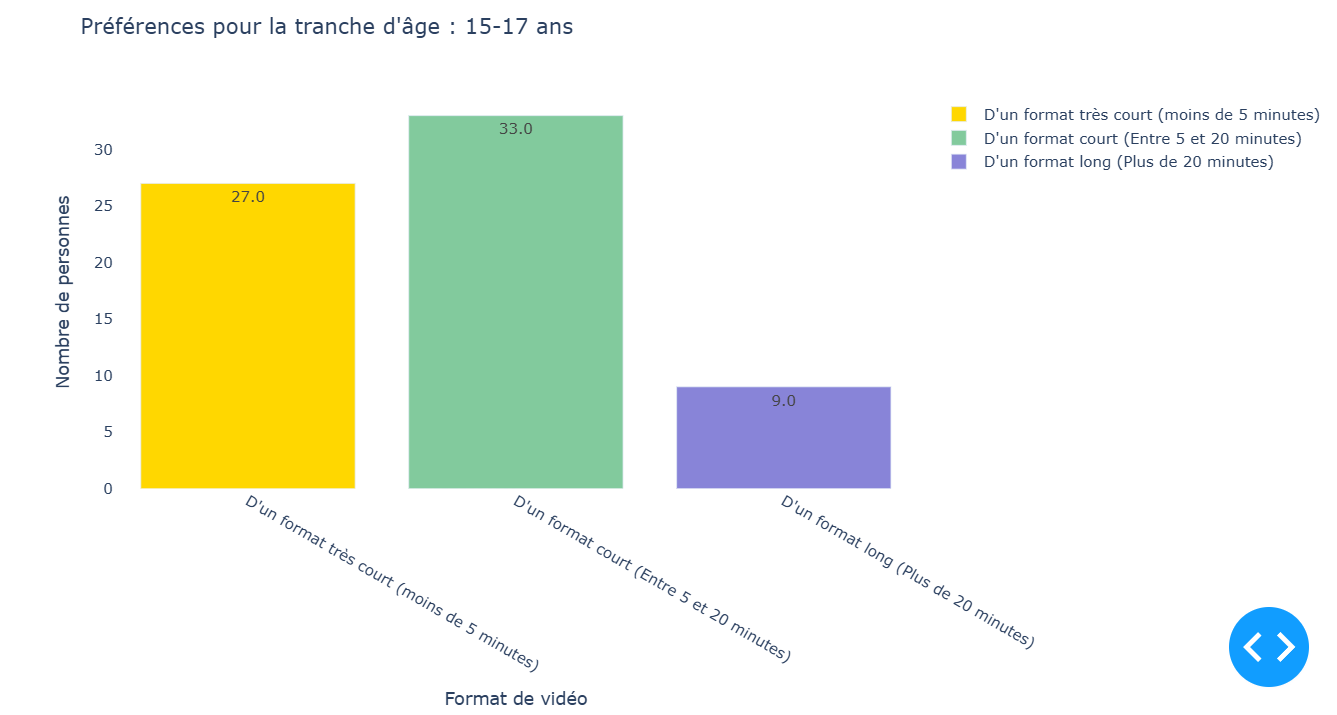
****

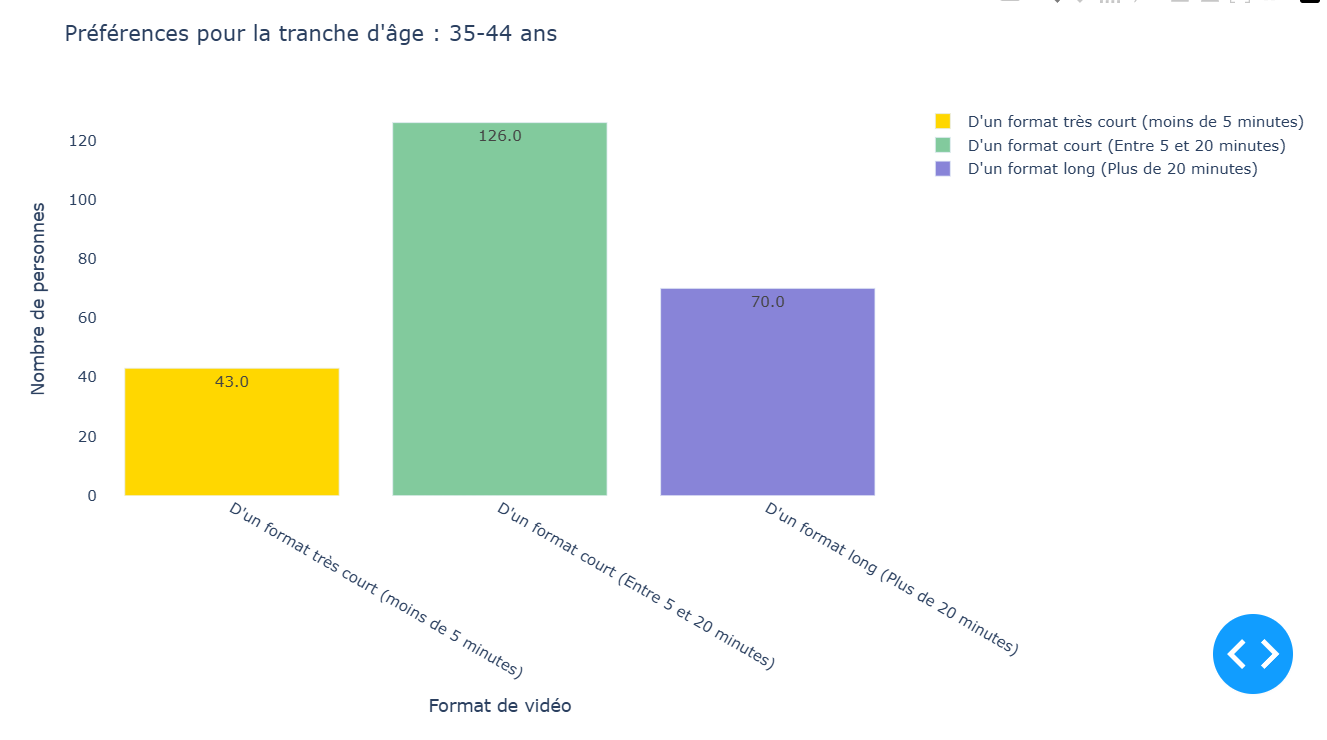
*Figure 38: Les fréquences des modes d’informations préférés pour chaque tranche d'âge*

****

*Figure 39: Les probabilités, pour chaque âge, qu’un mode d’information soit favori chez l’individu*

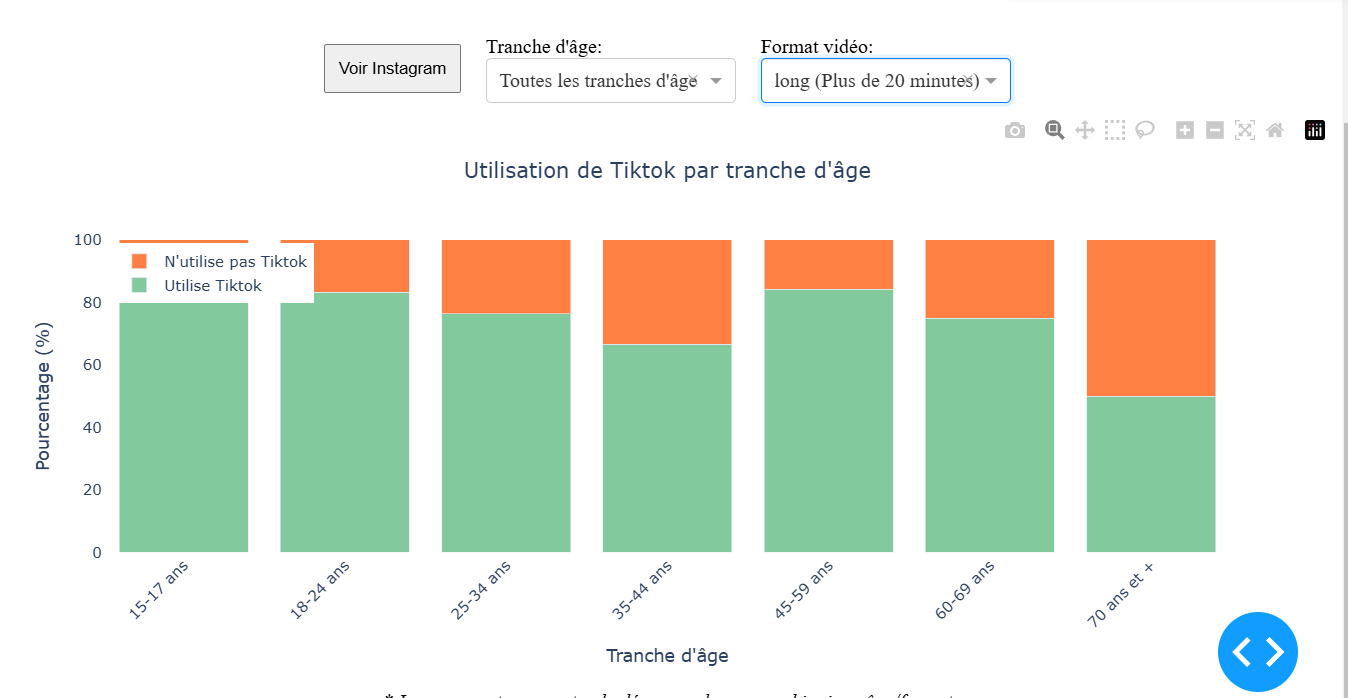
Les jeunes préfèrent donc les nouveaux médias, on a voulu vérifier par la suite si la consommation de ces nouveaux types de médias affectent la concentration des jeunes et l’attrait pour ceux-ci vis-à-vis des contenus très courts. Notre diagramme en histogramme semble confirmer nos théories. Toutes les générations semblent préférer en priorité un contenu court, puis ensuite long, et en dernier le contenu très court (dit “short”). Mais les jeunes générations ne partagent pas cette logique, en effet ils préfèrent le contenu court et le contenu court par rapport au contenu long.

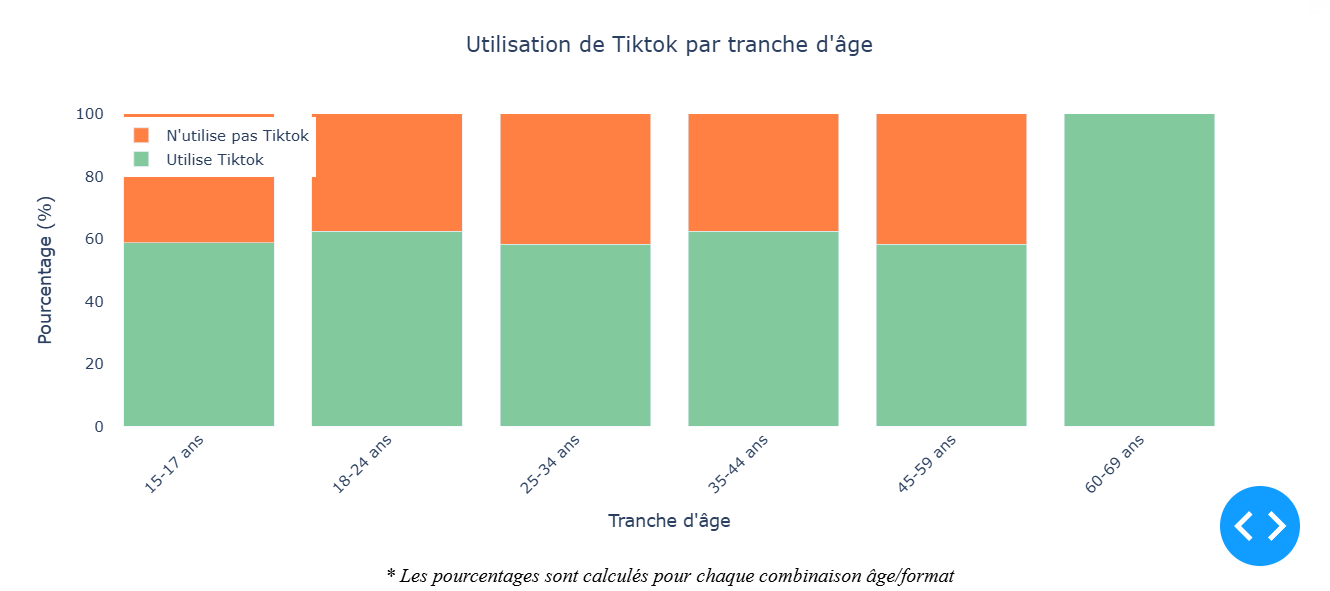




*Figure 40: Les préférences de format des 15-17 ans (en haut) et des 35-44 ans (en bas)*

Nous avons voulu mettre en relation ces données avec l’utilisation de Tiktok et Instagram (plateformes privilégiant le contenu court). Seulement, le problème est que notre dataset est trop faible, trop de personnes n’ont pas répondus à ces questions du sondage, et de ce fait, il semblerait que nos histogrammes soient incohérents par rapport aux résultats auxquels on s’attend



**

*Figure 41: L’utilisation de TikTok des gens préférant le contenu long (en haut) et préférant le contenu short (en bas) : des résultats incohérents*

Pour aller plus loin, nous aurions pu utiliser des datasets bien plus riches. Nous les avons d’ailleurs trouvé sur statista.com, ils sont payants mais nous avons pu contourner leur système afin d’avoir les jeux de données sur nos pc, mais dans un souci de légalité et de respect des lois française, nous avons préféré nous contenter de nos données et ne pas exploiter les données en question.

# Conclusion

## Conclusion

Pour rappel, le travail que nous avons réalisé se partageait au travers de plusieurs missions :

* La conception d’un dashboard concernant les thématiques abordées par plusieurs chaînes françaises lors des JT.
* La conception d’un dashboard traitant des temps de paroles des femmes à la télévision et à la radio.
* La conception d’un dashboard traitant

Cette expérience a été très enrichissante. En effet, la problématique du projet était stimulante et nous a permis de déceler des tendances à partir de données plutôt opaques. Nous estimons avoir réussi à proposer des tableaux de bords clairs et instinctifs, qui racontent une réelle histoire et valorisent énormément les données mises à disposition.

Nous avons eu la chance de pouvoir jouir de connaissances en data Visualisation de-par le parcours professionnel de certains membres de notre équipe, et bien que nous n’avons pas utilisé les logiciels dont nous sommes habitués, les bonnes pratiques de développement et de conception de dashboard se sont montrées très utiles à l’avancement rapide de nos tâches.

Les objectifs de travail ont aussi été remplis: chaque membre a bien eu un rôle clef à jouer parmi le groupe, et chacun d’entre nous est fier d’avoir contribué à sa manière à la réalisation de ce projet.

Les objectifs associés à nos missions semblent, selon nous, globalement atteints, dans le cadre proposé par le projet. En effet, le besoin d’adaptation et d’anticipation était au coeur de ce projet, et bien que de nombreux obstacles se soient présentés à nous, nous sommes convaincu d’avoir fait ce qu’il fallait pour les dépasser lorsque c’était possible, et pour les considérer dans nos réflexions lorsqu’ils étaient hors de notre portée.

Enfin, nous sommes satisfait de la collaboration que nous avons pu mener, notamment avec le groupe 7 principalement, qui nous a permis de cerner les enjeux d’une bonne communication, et d’accéder à une compréhension plus globale des enjeux de ce projet.

## Bilan

Ce projet s’est déroulé de la meilleure des façons: nous avons réussi à rapidement obtenir des visuels très satisfaisants, ce qui nous a mis en confiance et ce qui nous a permis de nous lancer rapidement dans des recherches de nouvelles données plus complètes.

Après un bilan de fin de première semaine, nous avons convenu tous ensemble de rapidement produire une première version des documents à rendre, afin de juger nos possibilités en termes d’exploration et d’aboutissement des problématiques dans le temps alloué au projet. Cette décision s’est avérée très avantageuse pour notre groupe, nous permettant de maîtriser au mieux notre timing et de ne pas subir la pression des derniers jours.

L’implication de l’ensemble des membres du groupe a été exemplaire tout au long du projet, ce qui a permis de conserver une bonne ambiance de travail du début à la fin. Nos choix dans la gestion du groupe, tel que la mise en place de présentations récapitulatives quotidiennes, a permis à l’ensemble des membres du groupe de se sentir intégré au projet, en plus d’améliorer leur compréhension de celui-ci, bien au-delà de leur travail personnel.

Le bilan de ce projet est pour nous positif, nous en ressortons satisfaits, et fiers de ne pas s’être laissé abattre par les manquements du côté des données. Cette expérience résonnera sans aucun doute dans nos projets futurs, comme preuve d’efficacité de la persévérance et de la bonne volonté face aux difficultés rencontrées.

# Sources

Catastrophes naturelles

<https://www.oise.gouv.fr/contenu/telechargement/14143/87201/file/Feuillet+n%C2%B0289+-+Catastrophes+naturelles++Bilan+2010.pdf>

Données audiovisuel

<https://defis.data.gouv.fr/defis/les-francaises-et-francais-face-a-linformation>

Données IMDB sur le genre au cinéma

[https://www.kaggle.com/code/jeremyverneuil/female-representation-in-movies-frenchinéma](https://www.kaggle.com/code/jeremyverneuil/female-representation-in-movies-frenchin%C3%A9ma)