PROJET RÉALISÉ PAR L'ÉQUIPE 4

INFLUENCE DE LA TV SUR LES ÉLECTIONS ÉLECTORALES

Aboubacar AMADOU, Hamza HAFSI, Anas JEBALI, Liam KHAMIS.

Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Mai 2023

Soumis comme contribution partielle
pour le cours Science des données 2 et Bases de données

Déclaration de non plagiat

Nous déclarons que ce rapport est le fruit de notre seul travail, à part lorsque cela est indiqué explicitement.

Nous acceptons que la personne évaluant ce rapport puisse, pour les besoins de cette évaluation:

- la reproduire et en fournir une copie à un autre membre de l'université; et/ou,
- en communiquer une copie à un service en ligne de détection de plagiat (qui pourra en retenir une copie pour les besoins d'évaluation future).

Nous certifions que nous avons lu et compris les règles ci-dessus.

En signant cette déclaration, nous acceptons ce qui précède.

Signature: AMADOU Aboubacar / 22002528 Date: 07/05/2023

Signature: HAFSI Hamza / 22003134 Date: 07/05/2023

Signature: JEBALI Anas / 22011379 Date: 07/05/2023

Signature: KHAMIS Liam / 22204118 Date: 07/05/2023

Table des matières

Chapitre 1

Introduction

1.1 Présentation du projet

Le processus électoral est une étape cruciale pour les démocraties du monde entier. Les élections présidentielles, en particulier, sont l'un des moments les plus importants dans la vie politique d'un pays. Le choix des électeurs est déterminé par divers facteurs, tels que les politiques, le charisme du candidat, son parcours, sa réputation et bien d'autres. Cependant, un autre facteur peut également jouer un rôle important dans la détermination du résultat des élections présidentielles : les médias.

Les médias ont une grande influence sur l'opinion publique, en particulier pendant les campagnes électorales, car ils sont souvent la principale source d'information pour les électeurs. La façon dont les médias couvrent les événements politiques, y compris les débats, les rassemblements électoraux et les interviews, peut avoir un impact significatif sur la perception qu'ont les électeurs des candidats.

A-t-on plus de chance d'être élu selon la chaîne de TV sur laquelle on passe ?

Pour tenter de répondre à cette problématique, nous nous munierons de plusieurs jeux de données, à savoir les différents temps de parole de chaque candidat sur chaque chaîne de TV [https://www.data.gouv.fr/fr/datasets/temps-de-parole-et-dant
les résultats du premier tour des élections présidentielles [https://www.data.gouv.fr/fr/datasets/election-presidentielle-des-10-et-24-avril-2022-resultats-defin
ainsi que les parts d'audience de chaque chaîne [https://www.data.gouv.fr/fr/datasets/audience-de-la-television].

Les résultats de cette étude pourront fournir des informations précieuses aux candidats, aux partis politiques et aux décideurs pour élaborer des stratégies efficaces de communication politique lors des prochaines élections présidentielles.

Chapitre 2

Base de données

2.1 Descriptif des tables

- La table Prise_parole : Cette table se compose d'id_prise qui est un integer unique permettant d'identifier chaque prise de parole individuelle, c'est aussi la clé primaire de la table. Ensuite, duree est un integer et correspond au temps de la prise de parole en minutes, pourcentage est un integer et correspond au pourcentage de chaque prise de parole par rapport au temps total autorisé à chaque chaîne de TV, nom_chaine est importée de la table Chaine et nom_candidat est importé de la table Candidat.
- La table Chaine : Cette table se compose de nom_chaine qui est un varchar unique et qui correspond au nom de la chaîne de TV et qui est la clé primaire de la table, tps_total est un integer qui correspond au nombre de minutes de temps de parole total de chaque chaîne, et part_audience est un integer qui correspond à la part d'audience de chaque chaîne.
- La table Candidat : Cette table se compose de nom_candidat qui est un varchar unique et correspond au nom et prénom de chaque candidat, c'est aussi la clé primaire de la table.
- La table Resultat : Cette table se compose d'id_resultat qui est un varchar unique et qui correspondpermet d'identifier chaque résultat, c'est aussi la clé primaire de la table. Ensuite nous avons departement qui est un varchar, voix_ex est un integer et correspond au nombre de voix exprimées pour chaque candidat. Les voix exprimées sont les voix des personnes qui ont voté pour un candidat valide, et non blanc. Enfin, nous avons pourcentage_voix qui est un decimal.

Le jeu de données de base (sur les temps de parole) comportant plus de 60 colonnes, la plupart ont été enlevées afin de ne gardder que celles qui étaient utiles à notre problématique. Pour ce qui est du jeu de données sur l'audience des chaînes de TV, nous avons trié les données de telle sorte à ne garder que les valeurs d'audience des chaînes présentes dans notre autre jeu de données.

2.2 Modèles MCD et MOD

- Une fois notre problématique et nos tables trouvées, nous avons réalisé le modèle conceptuel des données à l'aide de l'outil en ligne `Mocodo` [https://www.mocodo.net/]

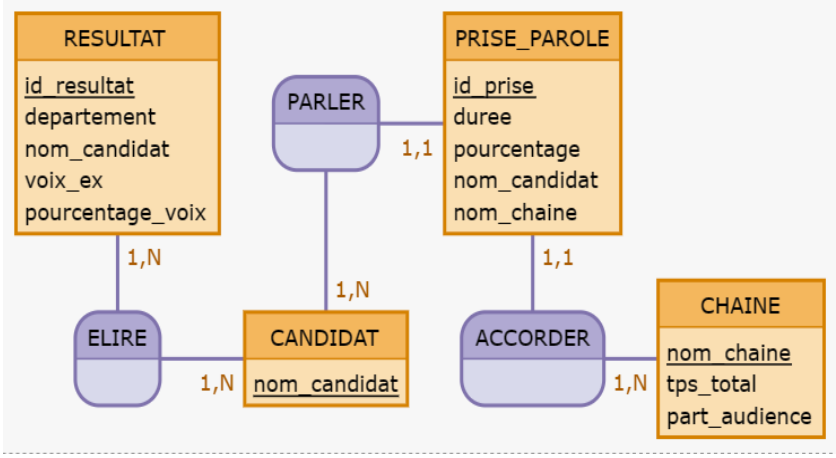


Figure 2.1: MCD.

- Pour le MOD, une version manuscrite a été réalisée en premier, puis la version issue du designer de phpMyAdmin une fois les données importées.

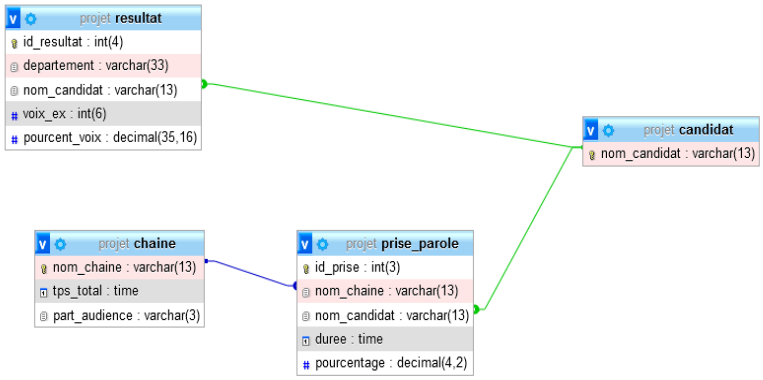


Figure 2.2: MOD.

Prise_parole (id_prise, duree, pourcentage, nom_chaine, nom_candidat)

Chaine (nom_chaine, tps_total, part_audience)

Candidat (nom_candidat)

Resultat (id_resultat, departement, voix_ex, nom_candidat, pourcentage_voix)

Elire (nom_candidat, departement)

Chapitre 3

Prétraitement des données

Dans un projet statistique, il semble évident que les données collectées ne suffisent pas à elles seules pour l'analyse. Les données brutes nécessitent souvent un prétraitement approfondi pour garantir leur exploitabilité. Notre groupe a consacré beaucoup de temps et d'efforts au prétraitement des données, notamment en supprimant les colonnes non pertinentes, en agrégeant certaines données entre elles. Ces étapes de prétraitement étaient essentielles pour rendre les données plus gérables et utilisables pour l'import en base de données SQL, l'analyse avec R.

3.1 Dataset des temps de parole

Comme dit précedemment, nous travaillé avec un ensemble de données contenant le temps de parole de chaque candidat sur différentes chaînes de télévision pendant la période électorale. Le jeu de données initial comportait plusieurs colonnes, dont certaines qui n'étaient pas pertinentes pour notre analyse, comme le type de temps de parole.

- Chaque ligne comporant "temps de parole total" a été conservée et les autres lignes ont été supprimées
- La colonne "Circonscription" ne contenait qu'une variable "NATIONAL" et était donc redondante, elle a été supprimée

3.2 Dataset des résultats

Nous avons été confronté à un défi de taille lors du traitement du deuxième jeu de données. En effet, avec plus de 60 colonnes et une ligne pour tout les candidats, l'ensemble de données était lourd et difficile à traiter. Cependant, nous avons utilisé des rêquetes "power-query" d'Excel pour simplifier le traitement des données. Les requêtes puissantes se sont révélées être un outil efficace pour simplifier les ensembles de données complexes, mais ont nécessité plusieurs heures de requêtes.

- Regroupement des bureaux de vote en départements et les lignes des candidats en une seule ligne pour chaque candidat
- Suppression d'énormément de colonnes inutiles telles que les votes blancs, % de votes blancs, nombre d'inscrits, nombre d'absents, ou encore numéro du bureau de vote

3.3 Dataset des audiences

Le troisième jeu de données contenait les audiences des 20 chaînes de télévision les plus importantes au cours des 40 dernières années, ainsi que d'autres feuilles Excel non utiles à notre projet. Cependant, notre projet ne nécessitait que les chaînes de télévision apparaissant dans nos autres jeux de données et leurs l'audience pour l'année 2020.

- Filtrage des chaînes de télévision non pertinentes et des données des autres années
- Suppression d'autres feuilles Excel non utiles à notre projet

Chapitre 4

Requêtes en langage naturel et SQL

4.1 Extraction du temps total d'antenne pour chaque candidat

Cette requête nous permet de récuperer le temps d'antenne total de chaque candidat en additionnant son temps sur chaque chaîne.

```
SELECT nom_candidat, sum(duree) as total_temps_de_parole
    FROM prise_parole
    GROUP BY nom_candidat
```

4.2 Extraction des candidats, des chaînes où chacun d'eux est passé et du nombre de voix total

Cette requête nous permet de récuperer le nombre total de voix de chaque candidat (en sommant celui dans chaque département) ainsi que la liste des chaînes sur lesquelles ce dernier est passé. On les tri ensuite par ordre décroissant du nombre de voix total

```
SELECT resultat.nom_candidat, prise_parole.nom_chaine, sum(voix_ex) as t
    FROM resultat, prise_parole, chaine
    WHERE prise_parole.nom_candidat = resultat.nom_candidat
    AND prise_parole.nom_chaine = chaine.nom_chaine
    GROUP BY nom_candidat, nom_chaine
    ORDER BY total_voix DESC
```

4.3 Extraction du temps d'antenne de chaque candidat sur chaque chaîne

Cette requête nous permet de récuperer le temps d'antenne de chaque candidat sur chaque chaîne.

```
SELECT nom_candidat, nom_chaine, SUM(duree) AS temps_antenne
    FROM prise_parole
    GROUP BY nom_candidat, nom_chaine
```

4.4 Extraction du nombre de voix de chaque candidat

Cette requête nous permet de récuperer le nombre de voix de chaque candidat (en sommant celui dans chaque département).

```
SELECT nom_candidat, SUM(voix_ex) AS nombre_de_voix
    FROM resultat
    GROUP BY nom_candidat
```

Chapitre 5

Analyse statistique

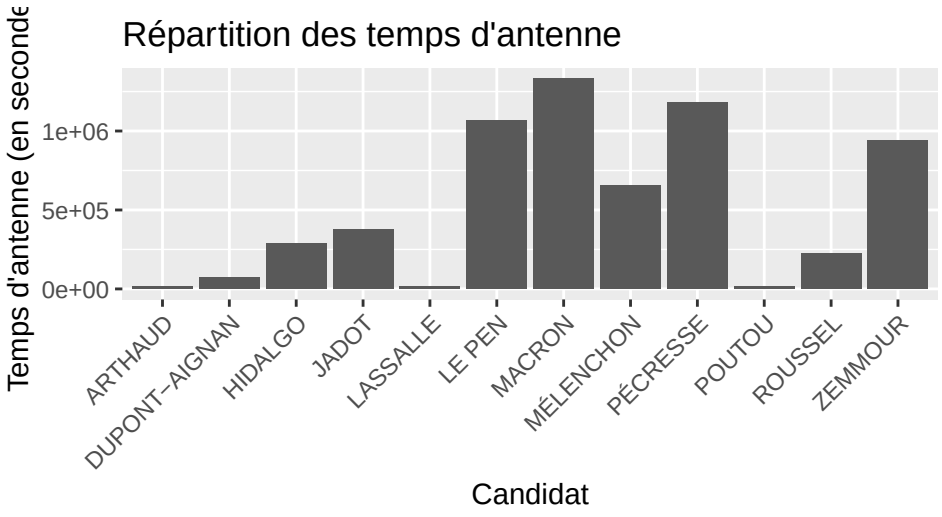
Maintenant que l'import des données est finalisé et que les requêtes ont été exécutées, nous pouvons passer à l'analyse statistique de ces données. Cette dernière sera composée de 3 sous parties; nous commencerons par **étudier la répartition des temps d'antenne des candidats sur les différentes chaînes de TV**, ensuite nous **analyserons l'impact des chaînes de TV sur les résultats électoraux**, pour enfin finir par **évaluer la corrélation entre le temps d'antenne sur chaque chaîne et le nombre de voix obtenues par chaque candidat.**

5.1 Etude de la répartition des temps d'antenne des candidats sur les différentes chaînes de télévision

Dans cette première sous-partie, nous allons étudier la répartition du temps d'antenne (en secondes) de chaque candidat à la télévision. Grâce à la requête 4.1, nous avons importé les données nécéssaires, nous pouvons donc créer un graphique à barre qui affiche le temps total de chaque candidat. Ce dernier est en secondes, afin de faciliter l'utilisation des données.

En utilisant la fonction ggplot, nous pouvons tracer ce graphique dans R :

```
ggplot(data1, aes(x = nom_candidat, y = total_temps_de_parole)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Répartition des temps d'antenne",
       x = "Candidat",
       y = "Temps d'antenne (en secondes)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



On peut constater d'après ce graphique que :

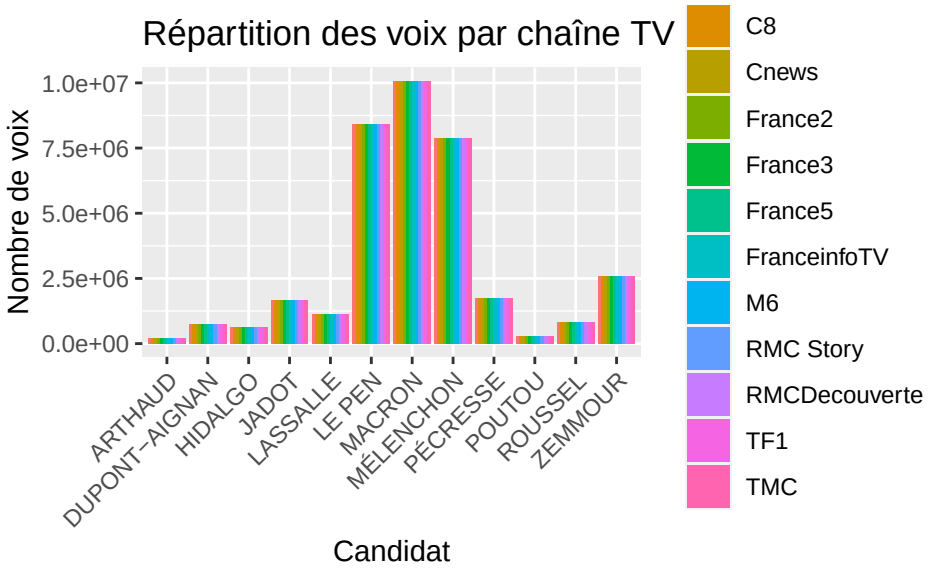
- Les candidats "MACRON", "PECRESSE", "LE PEN" et "ZEMMOUR" possèdent le plus de temps de parole sur les chaînes de télévision.
- Pour les autres candidats, on ne constate pas de différence significative entre leurs temps de parole totaux.

5.2 Analyse de l'impact des chaînes de TV sur les résultats électoraux

Dans cette partie, il sera question d'une analyse sur un possible lien entre la chaîne de télévision où les candidats ont obtenu leur audience et les voix obtenues à la suite de leur passage sur ces différentes chaînes. Pour cela, nous allons essayer de voir si chaque candidat est passé sur chaque chaîne de TV. Nous pouvons créer un graphique permettant de visualier le nombre de voix de chaque candidat, en changeant la couleur des barres afin de voir si chacun est passé sur chaque chaîne.

Le voici :

```
ggplot(data2, aes(x = nom_candidat, y = total_voix, fill = nom_chaine)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Répartition des voix par chaîne TV",
       x = "Candidat",
       y = "Nombre de voix",
       fill = "Chaîne TV") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Ce graphique montre que, malgré les différences de votes, tous les candidats sont passés sur toutes les chaînes de télévision.

5.3 Evaluation de la corrélation entre le temps d'antenne sur chaque chaîne et le nombre de voix obtenues par chaque candidat

Dans cette partie, il s'agira d'analyser l'impact de la durée d'antenne sur le nombre de voix obtenues par les candidats. Pour cela, nous avons extrait les données de la durée d'antenne et du nombre de voix pour chaque candidat à partir de la base de

données. Nous avons fusionné les deux tables de données et calculé la corrélation entre les deux variables.

Tout d'abord, nous avons calculé la corrélation entre les 2 variables à l'aide de la fonction cor :

```
cor(data5$temps_antenne, data5$nombre_de_voix)
```

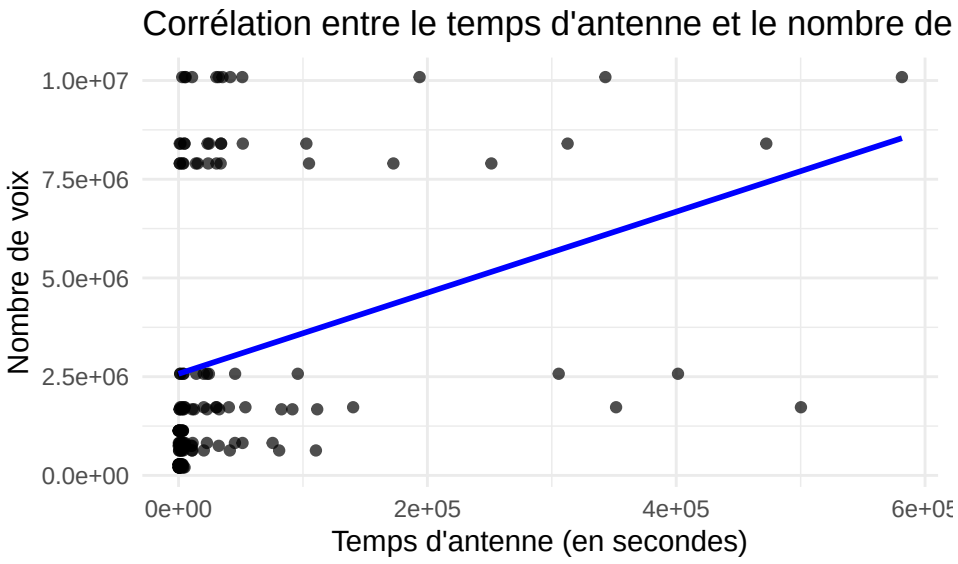
```
## [1] 0.2962079
```

Le résultat obtenu est environ de $0,30$.

A partir de cette corrélation, nous pouvons maintenant tracer en le graphique, obtenu grâce à une régression linéaire :

```
ggplot(data5, aes(x = temps_antenne, y = nombre_de_voix)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  ggtitle(paste("Corrélation entre le temps d'antenne et le nombre de voix (co
  xlab("Temps d'antenne (en secondes)") +
  ylab("Nombre de voix") +
  theme_minimal()
```

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



La corrélation étant faible, il semble qu'il n'y ai pas de lien significatif entre la durée d'antenne et le nombre de voix obtenues par les candidats.

Chapitre 6

Difficultés rencontrées

6.1 Choix des jeux de données supplémentaires

Dans le cadre de ce projet, le choix des données était d'autant plus compliqué car les jeux de données fournis ne couvraient pas l'ensemble des variables nécessaires à l'analyse. De plus, les données de médiamétries, qui auraient pu constituer une source d'informations importante, ne sont pas publiques et ne peuvent être facilement accessibles. Cela a rendu la tâche plus difficile en termes de collecte des données et nous a obligé à faire preuve de créativité pour trouver des sources de données alternatives. Le jeu de données supplémentaire étant lui même issue des données de médiamétrie, il a servi de passerelle avec les données privées pour récupérer les informations qu'il nous manquait.

6.2 Pré-traitement

Le prétraitement des données était une tâche difficile en raison de la manière dont les données étaient organisées. Les données brutes présentaient un grand nombre de colonnes, et tous les candidats étaient sur une seule ligne, rendant difficile l'extraction des informations nécessaires pour l'analyse. De plus, l'apprentissage de Power Query, l'outil utilisé pour le nettoyage des données, nécessitait une certaine expertise technique. Malgré les défis rencontrés, le prétraitement des données était une étape essentielle pour garantir l'exactitude et la fiabilité des résultats de l'analyse. De plus, cette étape a permit aux membres de l'équipe de gagner en expérience avec le logiciel Excel.

6.3 Hébergement de la base de données

L'utilisation d'une base de données hébergée en ligne était un problème car la solution d'hébergement gratuite que nous avons utilisée (filess.io) ne permettait pas plusieurs connexions simultanées. Par conséquent, chaque membre de l'équipe a dû héberger sa propre base de données, ce qui a réduit l'intérêt de la manœuvre collaborative, mais qui a permis à chaque membre de l'équipe d'apprendre à utiliser un service d'hébergement de base de données en ligne.

Chapitre 7

Conclusion et perspectives

Ce projet a été très intéressant et nous avons pu apprendre beaucoup de choses sur l'analyse de données avec R. Cependant, nous avons été limités par nos compétences en R et en statistiques. La partie sur l'analyse de l'impact des chaînes de télévision sur les résultats électoraux aurait pu être plus pertinente si nous avions évalué le pourcentage de chaque chaîne de télévision dans le temps d'antenne de chaque candidat, mais nous ne savions pas comment faire. Malgré cela, nous sommes satisfaits de ce que nous avons produit et de ce que nous avons appris. Ce projet peut certainement être approfondi et amélioré avec des compétences et des connaissances plus avancées.

Nous tenons à remercier nos professeurs pour leur aide tout au long de ce projet, malgré la période difficile que nous avons vécu. Il fût difficile de rester impliqué dans un projet totalement en distanciel, mais nos professeurs ont su nous motiver et nous aider lorsque nous en avions besoin.