Projet traitement de données massives : analyse et prétraitement des données

Etienne CANDAT, Lucie JEAN-LABADYE, Alice NOLLET

Abstract

Dans le cadre du cours "Analyse et traitement de données massives" de la session d'hiver 2022, nous devons conduire un projet ayant pour objectif de répondre à la question "Comment prédire la position politique d'un individu?" Nous nous appuierons sur les réponses fournies par plus de 37,000 électeurs canadiens à un sondage mené par le Consortium de la démocratie électorale avant et après l'élection fédérale de 2019. Ce premier rapport de projet fait état de nos réflexions concernant le jeu de données, son analyse et son prétraitement, ainsi que des pistes que nous pensons explorer pour le traitement et la prédiction.

1. Introduction

Le jeu de données constitué par le Consortium de la démocratie électorale contient plus de 23 millions d'entrées : il est par conséquent impossible de le soumettre tel quel à un quelconque modèle de prédiction. En plus de ses dimensions, la qualité des données récoltées, les modalités d'implémentation de la base de données et la forte présence de valeurs manquantes imposent un prétraitement rigoureux, dont nous ferons état ici.

Nous commencerons donc par détailler notre processus de prise en main du jeu de données, avant de décrire les quelques tests que nous avons conduits et de conclure sur les attributs que nous avons choisi de privilégier pour le premier temps de l'analyse. Nous évoquerons ensuite les méthodes envisagées pour mener cette dernière, et conclurons sur les procédures de tests nécessaires à la construction d'un modèle pertinent et efficace.

1.1. Dates importantes

Introduction au projet et mise à disposition de la base de données	12 janvier 2022
Rendu du premier rapport	16 février 2022
Rendu du rapport final et des résultats	13 avril 2022

Table 1. Dates importantes

2. Analyse des données et de leurs propriétés statistiques, et premières réflexions sur les attributs initiaux

Avant même de commencer à manipuler la base de données, nous nous sommes intéressés au PDF de documentation fourni, dont nous avons naturellement commencé la lecture par l'introduction. La partie 5 de cette dernière, "Sample design and survey weights" a particulièrement attiré notre attention, car elle informe sur la façon dont les données ont été collectées, et l'échantillon constitué. Nous y avons appris deux choses importantes. Premièrement, il est expliqué que les données ont été récoltées en deux temps : 37,822 personnes ont été interrogées avant l'élection, constituant la partie "CPS" de la base de données, et 10,340 d'entre elles (environ 27 % des répondants initiaux) ont été recontactées après l'élection, afin de répondre à un questionnaire de suivi. Leurs réponses sont enregistrées sous le préfixe "PES". Nous nous sommes alors interrogés sur la répartition des données

"CPS" et "PES", et elle s'est avérée relativement équilibrée : 41% des attributs commencent par "CPS", et 47% d'entre eux sont siglés "PES". Si cet équilibre relatif nous a d'abord amenés à penser qu'il serait judicieux de se concentrer sur les attributs "CPS", car moins nombreux et (nous le pensions), similaires aux observations "PES", il peut toutefois engendrer d'autres déséquilibres, que nous mentionneront plus tard. Dans un second temps, le paragraphe nous informe de la façon dont a été constitué l'échantillon des répondants. L'équipe y renseigne qu'elle a tenté de respecter la parite femmes/hommes, et qu'elle s'est surtout concentrée sur les tranches d'âge, les quotats provinciaux et la langue principale de l'échantillon. Ici aussi, ces préférences peuvent impliquer le déséquilibre d'autres variables qui peuvent s'avérer significatives, comme la religion, l'activité ou la catégorie salariale. Pour donner un exemple, la médiane de la colonne "cps19_income_number"[1] est des 72,000 dollars, alors que le salaire médian des ménages canadiens est annoncé à 62,900 dollars en 2018, d'après Statistique Canada[2]. Nous n'excluons pas l'influence de facteurs extérieurs pour expliquer cette différence, mais c'est une limite de nos données que nous avons voulu inerroger. Enfin, l'introduction nous renseigne sur la qualité des données et les métadonnées incluses à hauteur de 11% dans nos variables.

Pour la suite du document, nous avons divisé la lecture, afin de se faire une idée claire et exhaustive des attributs disponibles. Chaque membre de l'équipe devait alors prendre note de ceux qui lui semblaient pertinent, et leur attribuer un poids. Lors de notre mise en commun, nous nous sommes rendu compte qu'ils pouvaient être divisés très grossièrement en deux catégories :

- Les attributs que nous avons identifés comme "factuels", descriptifs (l'âge, le genre, le niveau d'éducation, etc.), qui représentent 39% des observations "CPS" et 14% des attributs "PES",
- Les attributs que nous avons appelés "d'opinion", dont la question associée commence généralement par "Pensez-vous que...". Ils représentent donc environ 50% des "CPS" et 75% des PES, en prenant en compte les métadonnées.

Or, il s'est avéré qu'une grande partie des attributs auxquels nous avions attribué des poids importants relevaient de la catégorie "d'opinion", ce qui a fortement questionné notre choix de nous concentrer sur les attributs "CPS", ces derniers étant plus souvent "factuels" que les "PES".

Enfin, après avoir posé ces premières réflexions et avoir lu la documentation, nous avons commencé à prendre en main notre jeu de données. Nous avons commencé par évaluer ses dimensions, et avons vite compris que nous allions faire face au fléau de la dimensionnalité : si nous prenons comme référence l'article "Optimal number of features as a function of sample size for various classification rules"[3], le ratio idéal pour un jeu contenant des données corrélées de $\sqrt{37822} \approx 194$ variables est 3 fois dépassé, et à cela s'ajoute le grand nombre de valeurs manquantes, qui représentent en moyenne 54% du contenu d'une observation. Il est donc évident qu'il sera nécessaire de réduire les dimensions avant de pouvoir déployer un quelconque modèle.

Une autre difficulté s'est présentée au moment d'étudier la structure du jeu de données, en particulier le typage de ses variables : 78% d'entre elles sont de type "object", plus complexe à traiter que le type numérique. De plus, au moins 6% de ces variables correspondent à des zones de textes libres, qui doivent elles aussi faire l'objet d'un prétraitement particulier (gestion de la casse, des coquilles, etc.), qui peut s'apparenter à de la gestion de bruit. Nous supposons par ailleurs une faible présence de ce phénomène, les données n'étant pas captées

[1] Voir le lexique en annexe, et ce pour toutes les occurences de nom d'attribut

[2] URL : https://www150.statcan.gc.ca/t1/tbl1/fr/tv.action?pid=1110023901

[3] Jianping Hua, Zixiang Xiong, James Lowey, Edward Suh, Edward R. Dougherty. 2005. "Optimal number of features as a function of sample size for various classification rules." Bioinformatics. Volume 21 (issue 8), pp. 1509-1515

par un système sujet à erreur, mais resterons attentifs aux possibilités de réponses trop hâtives ou de répondants n'ayant pas pris le sondage au sérieux, que la documentation sur la qualité des données évoque.

3. **Choix des attributs initiaux**

Face à la masse de données que nous avons à notre disposition, un choix cohérent et pertinents d'attributs initiaux constitue une bonne première étape dans la réduction des dimensions. Comme évoqué dans la partie 2, nous avons tout de suite étudié le jeu de données par le prisme de cette sélection.

Après la lecture complète de la documentation, nous avons échangé sur les attributs auxquels nous avons donné les poids les plus importants, et avons remarqué que nous avions tous choisi des attributs dont le format de réponse se rapprochait du format de notre variable cible "cps19_votechoice", soit une liste de partis, ou au moins d'opinions partisanes. Ont ainsi été retenus les attributs suivants :

- cps19_imp_loc_iss_p et cps19_imp_iss_party
- cps19_v_advance, cps19_vote_lean
- cps19_2nd_choice
- cps19_not_vote_for
- cps19_outcome_most et cps_19_outcome_least
- cps19_fed_id et cps19_prov_id
- cps19_vote_2015
- pes19_party_rep_whic
- pes19_votechoice2019
- pes19_provvote
- pes19_pid et pes19_pidtrad

Nous avons ensuite échangé sur les attributs auxquels nous avions attribué un poids plus faible juste après ceux cités ci-dessus, et avons noté qu'ils étaient nettement moins homogènes : si certains présentaient eux aussi une liste des partis, ils allaient être plus difficiles à traiter car leur format n'était pas une simple liste (comme la matrice de cps19_issue_handle), tandis que d'autres faisaient référence au nom de candidats (les cps19_lead_int par exemple), et d'autres enfin relevaient d'une opinion dépassant le cadre partisan. Tous les attributs de cette première sélection avaient toutefois un point commun : ils pouvaient tous être considérés comme ce que nous avons appelé à la partie 2, des attributs "d'opinion".

Nous avons alors questionné notre tendance à laisser de côté les attributs "factuels", et avons décidé de conduire quelques tests simples sur ceux-ci afin de déterminer si nos premières intuitions étaient les bonnes. Nous avons alors décidé de se concentrer sur les 6 attributs utilisés par le Consortium de la démocratie électorale pour identifier les personnes ayant répondu plusieurs fois au questionnaire, à savoir :

- cps19_yob
- cps19_gender
- cps19_education
- cps19_employment
- cps19_religion
- cps19_bornin_canada

et ce pour deux raisons : premièrement car le Consortium les avait identifiées comme suffisamment "existentielles" pour identifier un individu unique, mais aussi et surtout car ces variables ne contiennent aucune donnée manquante.

Les tests réalisés étaient très simples et consistaient en une matrice de contingence entre la variable cible cps19_votechoice et l'attribut sélectionné. La plupart d'entre eux (à

l'exception de cps19_gender et cps19_bornin_canada) ont toutefois demandé un prétraitement, les données étant difficilement lisibles en l'état. Ainsi, nous avons ajouté au jeu de données les classes de valeurs suivantes (sous R) :

- Pour cps19_yob, nous avons reconstitué les classes d'âge que le Consortium évoque en introduction, à savoir 18-34 ans, 35-54 ans et 55 ans et plus. Nous retenons de la matrice de répartition des âges au sein des partis[4] que plus de la moitié des électeurs du Bloc Québecois ont plus de 55 ans, et qu'environ un tiers des électeurs des partis moins traditionnels que sont le NDP, le Green Part et le People's Party ont entre 18 et 24 ans. La matrice de répartition des votes par classe d'âge[5] quant à elle nous permet d'émettre les hypothèses selon lesquels la classe des 55 ans et plus a tendance à s'orienter vers les Conservateurs, et que de façon générale, leurs votes sont centralisées autour des 2 grands partis traditionnels, tandis que les 18-24 se dispersent un peu plus, et s'orientent plus sur la gauche de l'échiquier politique.
- Pour cps19_education, nous avons grossièrement divisé le groupe entre les personnes ayant arrêté leur scolarité avant l'université et les personnes étant entrées à l'université (ayant validé une ou plusieurs années ou non). Les graphiques en secteurs obtenus [6] tendent à mettre en évidence que les répondants étant allés à l'université ont plus tendance à s'orienter vers les partis libéraux (ils sont d'ailleurs les électeurs majoritaires du Parti Libéral), mais il faut garder en tête que la division grossière de 11 variables en 2 sous-groupes nous permet de n'émettre que des hypothèses.
- Pout cps19_employment, nous avons divisé tout aussi grossièrement entre les personnes ayant une activité et les personnes n'en ayant pas (comprenant ainsi aussi bien les personnes sans-emploi que les étudiants ou les personnes s'occupant de familles non-salariés). Nous notons ici que cette division grossière peut amener à estimer que le statut d'emploi d'une personne n'est pas significatif, et relève lui-même de variables explicatives comme l'âge, le genre ou le niveau d'éducation, l'analyse devant être affinée pour l'établir sérieusement.
- Pour cps19_religion, nous avons d'abord découpé la base en grands courants religieux (chrétiens, musulmans, juifs, religions "asiatiques" et non-religieux) puis nous avons subdivisé le sous groupe "chrétiens" en plus petites divisions correspondant aux différentes branches du christianisme. Ce cloisonnement ne s'est pas avéré très significatif, si ce n'est peut-être pour affiner les velléités conservatrices de quelqu'un se définissant comme chrétien. En ce qui concerne la division en courant religieux, la matrice de contingence obtenue [7] a peut-être été la plus significative dans ce qu'elle renferme d'information, nous apprenant que 61% des personnes se définissant musulmanes se reconnaissent dans le Parti Libéral, valeur de contingence non-négligeable. Les 2 autres grands monothéïsmes et les non-religieux se divisent quant à eux plutôt équitablement entre les 2 partis traditionnels, qui encore une fois, centralisent les intentions de votes ; tandis que les hindous, boudhistes et sikhs semblent préférer le Parti Libéral.

En ce qui concerne l'attribut cps19_gender[8], il peut être considéré comme significatif dans la mesure où les femmes ont tendance à moins s'orienter vers les Conservateurs, sans pour autant grossir beaucoup plus l'électorat d'autres partis, à part peut-être celui du NDP, la différence remarquable se situant dans le secteur "Don't know / Prefer not to answer".

[4] Annexe 2 - Figure 1
[5] Annexe 2 - Figure 2
[6] Annexe 2 - Figures 3 à 6
[7] Annexe 2 - Figure 7
[8] Annexe 2 - Figures 8 et 9

Une hypothèse que l'on peut avancer est le fait que les femmes se sentent moins concernées par la politique, qui est un monde d'hommes, ou ont plus de mal à formuler une intention affirmée. Cette hypothèse doit toutefois faire l'objet d'analyses plus poussées avant d'être sérieusement prise en compte.

Enfin, l'attribut cps19_bornin_canada ne s'est pas révélé significatif, les répartitions étant équilibrées entre les canadiens de naissance et les répondants rendus citoyens ou résidents permanents.

Nos premières inuitions et notre volonté de ne se concentrer que sur des attributs "d'opinion" se sont donc trouvées lacunaires. Si 2 attributs parmi les 6 semblent peu significatifs, ou devant faire l'objet d'une exploration plus approfondie pour le devenir, le reste des hypothèses évoquées concernent 75% de la population sondée, 25% du contenu de cps19_votechoice étant manquant, proportion non-négligeable et inatteignable pour les attributs listés au début de cette section, qui contiennent une valeur médiane de 19237.5 valeurs manquantes. Ils feront toutefois eux aussi l'objet d'une étude attentive, et si le temps nous le permet, nous incorporerons après eux les attributs de poids moindre évoqués plus tôt.

4. **Traitement pratique des données**

Si les choix d'attributs initiaux évoqués précédemment peuvent être considérés comme un premier élaguage de la base, sa dimmensionalité impose un traitement plus sérieux, celui-ci devant concerner à la fois la forme du jeu de données, et le fond des informations qu'il contient. Pour la forme, nous avons pensé fusionner les variables que l'on a jugées pertinentes, mais redondantes (décrivant un même fait), ou dont la mise en forme peut-être raccourcie (nous pensons ici aux questions ayant le même énoncé, mais des modalités de réponses différentes, commes pes19_partic$X$, ou cps19_lead_int, que l'on pourrait résumer en matrice de booléens pour faciliter le traitement), limitant ainsi la démultiplication.

Pour ce qui est du fond, nous avons envisagé plusieurs méthodes, que nous mettrons en place en fonction des types de variables que nous aurons à manipuler et/ou des sorties attendues. La plupart des méthodes présentées ci-après présentent toutefois un incovénient de taille, dans la mesure où leur déploiement recquiert un jeu de données exempt de valeurs manquantes. La gestion de ces dernières sera sans doute un des grands enjeux de ce projet, auquel nous avons décidé de faire face à l'aide des deux grandes méthodes les plus courantes. La première consiste à imputer la valeur de ces variables à l'aide d'outils statistiques plus ou moins complexes, tandis que la deuxième consiste en leur suppression pure et simple, en prenant néanmoins en compte leur signifiance et le taux de valeurs manquantes.

Une fois la base nettoyée, nous pensons mettre en place différents algorithmes de sélection des données, comme l'élimination descendante (backward elimination), qui nous permettra de ne conserver que les variables ayant un taux de contribution supérieur à un p que l'on se fixera (on optera certainement pour le seuil prudent $p = 0.1$). La prépondérance du type "object", qui caractérise plus de 75% des données diponibles, nous a aussi amenés à nous intéresser au test du Chi-2, ainsi qu'à diverses techniques de NLP comme la reconnaissance de patrons fréquents, qui pourrait par exemple nous permettre de réduire les entrées libres à leurs mots-clé.

Lorsque la réduction de dimension sera bien avancée, nous pensons déployer différents modèles afin de produire la prédiction finale : nous avons d'abord pensé mettre en place des méthodes de clustering, comme le partitionnement en k-moyennes, qui a l'avantage de pouvoir être déployé assez tôt pour explorer les données, comme beaucoup plus tard, pour catégoriser les variables à prédire. Nos réflexions se sont aussi tournées vers la régression logistique binomiale (tel répondant va-t-il voter pour tel parti?) ou multinomiale (prédiction directe de l'intention de vote), ainsi que vers les arbres de décision, qui ont le mérite d'être plutôt simples à mettre en place, mais l'énorme inconvénient d'être assez peu robustes si une

erreur se glisse proche de la racine. La plupart de ces méthodes seront implémentées à l'aide de la librairie Scikit-Learn, qui offre des outils puissants et optimisés pour leur déploiement.

5. Procédure de test envisagée

Afin de s'assurer de la pertinence et de l'efficacité des méthodes évoquées ci-dessus, il sera nécessaire de conduire différents tests tout au long du projet. La procédure qui nous a semblé la plus pertinente est la validation croisée k-échantillons, qui permet d'éviter le surapprentissage et l'overfitting, en gardant en tête que les déséquilibres de notre jeu de données nous obligeront certainement à le stratifier.

Appendix A. Lexique des noms d'attributs mentionnés dans leur ordre d'apparition

cps19_income_number répond à la question "Quel est le revenu total de votre ménage avant impôts en 2018? Cela doit inclure toutes les sources de revenus au millier de dollars près."

cps19_votechoice répond à la question "Pour quel parti prévoyez-vous de voter?"

cps19_imp_loc_iss_p répond à la question "Quel parti aborde le mieux cet enjeu local?"

cps19_imp_iss_party répond à la question "Quel parti aborde le mieux cet enjeu?"

cps19_v_advance répond à la question "Pour quel parti avez-vous voté ?"

cps19_vote_lean répond à la question "Etes-vous tenté(e) d'appuyer un parti en particulier ?"

cps19_2nd_choice répond à la question "Et quel parti serait votre second choix ?"

cps19_not_vote_for répond à la question "Y a-t-il un ou des partis pour lesquels vous ne voteriez absolument pas? (Veuillez sélectionner tous les partis qui s'appliquent)"

cps19_outcome_most répond à la question "Quel résultat électoral préféreriez-vous le plus?"

cps19_outcome_least répond à la question "Quel résultat électoral préféreriez-vous le moins?"

cps19_fed_id répond à la question "En politique fédérale, vous considérez-vous habituellement comme étant :"

cps19_prov_id répond à la question "En politique provinciale, vous considérez-vous habituellement comme étant:"

cps19_vote_2015 répond à la question "Pour quel parti avez-vous voté? (lors de l'élection fédérale de 2015)"

pes19_party_rep_whic répond à la question "Quel parti représente le mieux vos points de vue?"

pes19_votechoice2019 répond à la question "Pour quel parti avez-vous voté? (question post-élection)"

pes19_provvote répond à la question "En politique provinciale, vous considérez-vous habituellement comme étant:"

pes19_pid répond à la question "De quel parti vous sentez-vous le plus proche?"

pes19_pidtrad répond à la question "En politique fédérale, vous considérez-vous habituellement comme étant :"

cps19_issue_handle répond à la question "Quel parti aborderait le mieux chacun de ces enjeux?"

cps19_lead_int répond à la question "Parmi les chefs de partis fédéraux énumérés ci-dessous, le(s)quel(s) trouvez-vous intelligent(s)? (Sélectionnez tous ceux qui s'appliquent)"

cps19_yob répond à la question "Tout d'abord, en quelle année êtes-vous né(e)?"

cps19_gender répond à la question "Êtes-vous... (un homme/une femme/autre)"

cps19_education répond à la question "Quel est votre plus haut niveau de scolarité complété?"

cps19_employment répond à la question "Quel est votre statut d'emploi actuel?"

cps19_religion répond à la question "Quelle est votre religion, si vous en avez une?"

cps19_bornin_canada répond à la question "êtes-vous né au Canada?"

Appendix B. Tableaux et graphiques

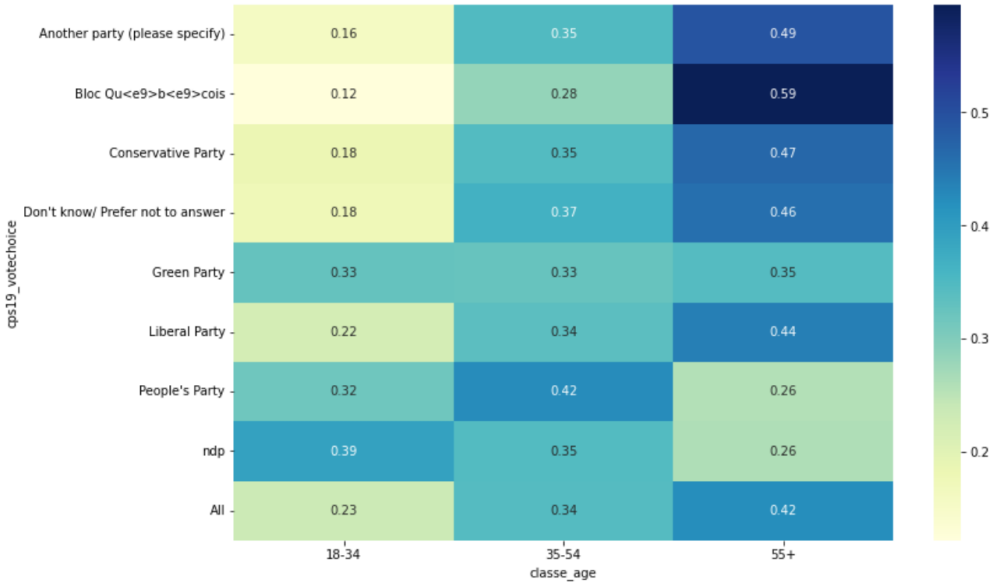

Figure 1. Matrice de contingence de la répartition des classes d'âges au sein des partis

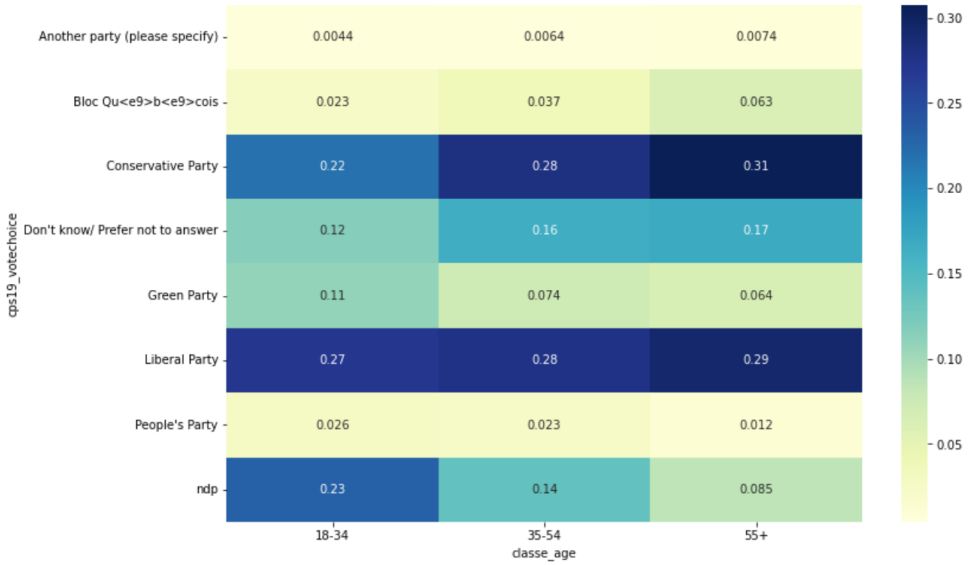

Figure 2. Matrice de contingence de la répartition des votes au sein des classes d'âges

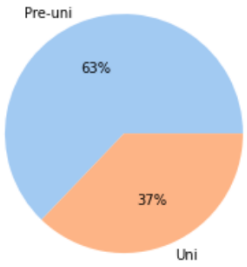

Figure 3. Répartition des électeurs du People's Party en fonction de leur niveau d'éducation

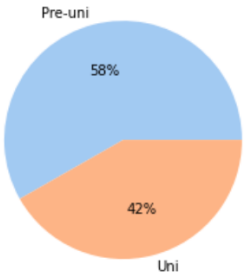

Figure 4. Répartition des électeurs du Parti Conservateur en fonction de leur niveau d'éducation

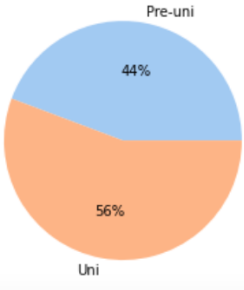

Figure 5. Répartition des électeurs du Parti Libéral en fonction de leur niveau d'éducation

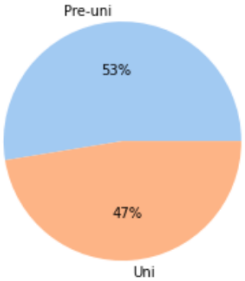

Figure 6. Répartition des électeurs du NDP en fonction de leur niveau d'éducation

Figure 7. Répartition des votes en par appartenance religieuse

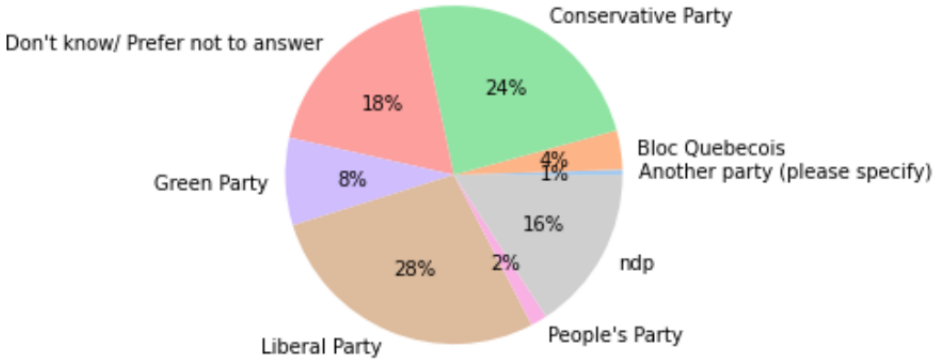

Figure 8. Répartition du vote des femmes

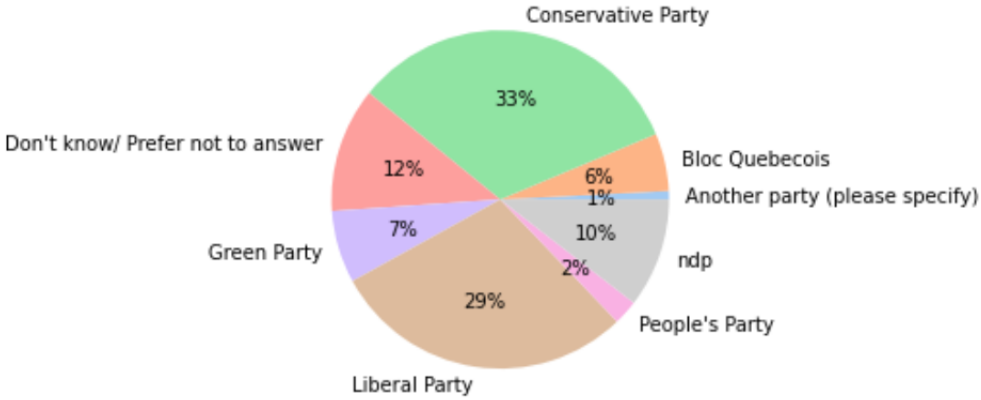

Figure 9. Répartition du vote des hommes