Présentation des variables

José Manuel Rodríguez Caballero

STT-7335 Méthodes d'analyse de données

March 3, 2025

Contents

1 Description du jeu de données

Le présent rapport s'appuie sur un ensemble de 3443 vers, chacun annoté par un ensemble de variables descriptives concernant l'auteur ou l'autrice, ainsi que par plusieurs scores émotionnels.

1.1 Détails des variables du jeu de données

La liste ci-dessous décrit les colonnes du fichier source :

- **poet_id** : Identifiant unique associé à chaque poète.
- **poet** : Nom du poète ou de la poétesse.
- **suicidal** : Indicateur booléen précisant si l'auteur ou l'autrice s'est suicidé(e). Peut prendre la valeur `TRUE` ou `FALSE`.
- **period** : Période d'écriture (par exemple, *Early* ou *Modern*).
- **sex** : Sexe de la personne à l'origine du poème (`Male` ou `Female`).
- **heterosexual** : Indicateur booléen spécifiant l'orientation sexuelle, `TRUE` pour hétérosexuel(le), `FALSE` dans les autres cas.
- **date_of_birth** : Date de naissance de l'auteur ou de l'autrice.
- **date_of_death** : Date de décès de l'auteur ou de l'autrice, si disponible.
- **country_of_birth** : Pays de naissance de l'auteur ou de l'autrice.
- **poem_title** : Titre du poème auquel le vers appartient.
- **n_words** : Nombre de mots contenus dans le vers.
- **anger** : Score reflétant la présence de la colère dans le vers (valeur numérique).
- **anticipation** : Score reflétant l'anticipation (ou l'attente) véhiculée dans le vers (variable réponse principale de cette étude).
- **disgust** : Score reflétant la présence du dégoût dans le vers.
- **fear** : Score reflétant la présence de la peur dans le vers.
- **joy** : Score reflétant la présence de la joie dans le vers.
- **sadness** : Score reflétant la présence de la tristesse dans le vers.
- **surprise** : Score reflétant la présence de la surprise dans le vers.
- **trust** : Score reflétant la confiance véhiculée par le vers.
- **negative** : Score global regroupant des émotions à valence négative.
- **positive** : Score global regroupant des émotions à valence positive.
- **verse** : Position normalisée du vers dans le poème (`0` : début, `1` : fin).

Après nettoyage et vérification de la cohérence des données, le corpus final contient 3443 vers annotés. Le nombre de données manquantes est faible et n'impacte pas significativement les analyses présentées ci-après.

1.2 Provenance des données

Les poèmes analysés dans cette étude ont été recueillis à partir de plusieurs sites Web, dont les références figurent dans les données brutes. Chaque poème a été divisé en vers, malgré un risque d'imperfection dans ce processus : un petit nombre de lignes ne contiennent par exemple que des numéros ou des citations placées avant le début réel du poème. Pour extraire les différentes composantes émotionnelles de chaque vers, nous avons recouru à la bibliothèque *syuzhet* du logiciel R. Celle-ci calcule divers scores émotionnels (colère, joie, peur, etc.) et facilite ainsi l'analyse lexicale et sentimentale des textes poétiques. Les dates de naissance, de décès et le pays de naissance des poètes proviennent de Wikipédia. Enfin, la classification des poèmes selon la période de l'auteur (début, milieu et fin) a été établie à l'aide de ChatGPT o1 pro.

1.3 Orientation sexuelle

Pour déterminer quels poètes sont hétérosexuels, nous avons utilisé le prompt suivant :

```
Génère un fichier CSV dont la première colonne s'intitule poet et la
deuxième colonne s'intitule heterosexual.

La colonne poet doit inclure les noms suivants :
- Adrienne Rich
- Alfred Edward Housman
- Anne Sexton
- Charlotte Mew
- Denise Levertov
- Edith Sitwell
- Edna St. Vincent Millay
- Hart Crane
- John Berryman
- John Davidson
- Lawrence Ferlinghetti
- Randall Jarrell
- Robert Lowell
- Sara Teasdale
- Sylvia Plath
```

```
- William Carlos Williams

La colonne heterosexual doit prendre la valeur 1 si le poète est
hétérosexuel, 0 s'il est homosexuel ou bisexuel, et NA si ces
informations ne sont pas clairement corroborées.

Merci de fournir simplement le tableau au format CSV, sans informations
superflues.
```

Ce prompt a été appliqué à ChatGPT o1 pro.

1.4 Données manquantes

Parmi les 16 poètes étudiés, un seul n'a pas pu être classé selon son orientation sexuelle. Nous projetons de récupérer cette donnée manquante en recourant à une analyse discriminante linéaire, qui permettra de prédire l'orientation sexuelle en fonction des émotions et d'autres variables.

1.5 Objectif

L'objectif de l'étude est d'évaluer dans quelle mesure les variables explicatives (par exemple, **anger**, **fear**, etc.) influencent la variable de réponse **anticipation**.

2 Présentation des variables

2.1 Statistiques descriptives

Le tableau ci-dessous présente un résumé statistique (minimum, quartiles, médiane, moyenne, maximum) des variables numériques, parmi lesquelles **anticipation**, **colère**, **peur**, etc.

Table 1: Variables numériques : statistiques récapitulatives

Variable	Min	1er quartile	Médiane	Moyenne	3e quartile	Max
Anticipation	0	0.00	0.0	0.15	0.00	3
Colère	0	0.00	0.0	0.11	0.00	4
Confiance	0	0.00	0.0	0.13	0.00	3
Dégoût	0	0.00	0.0	0.12	0.00	2
Joie	0	0.00	0.0	0.14	0.00	3
Nombre de mots	1	5.00	7.0	6.76	8.00	40
Négatif	0	0.00	0.0	0.33	1.00	4
Peur	0	0.00	0.0	0.17	0.00	4
Positif	0	0.00	0.0	0.27	0.00	4
Surprise	0	0.00	0.0	0.08	0.00	2
Tristesse	0	0.00	0.0	0.19	0.00	3
Vers	0	0.24	0.5	0.49	0.74	1

2.2 Distribution de la variable réponse

La Figure 1 illustre la distribution de la variable **anticipation** sous forme d'histogramme. On constate une forte concentration des valeurs à 0 ainsi que quelques occurrences plus élevées, bien que rares dans le corpus.

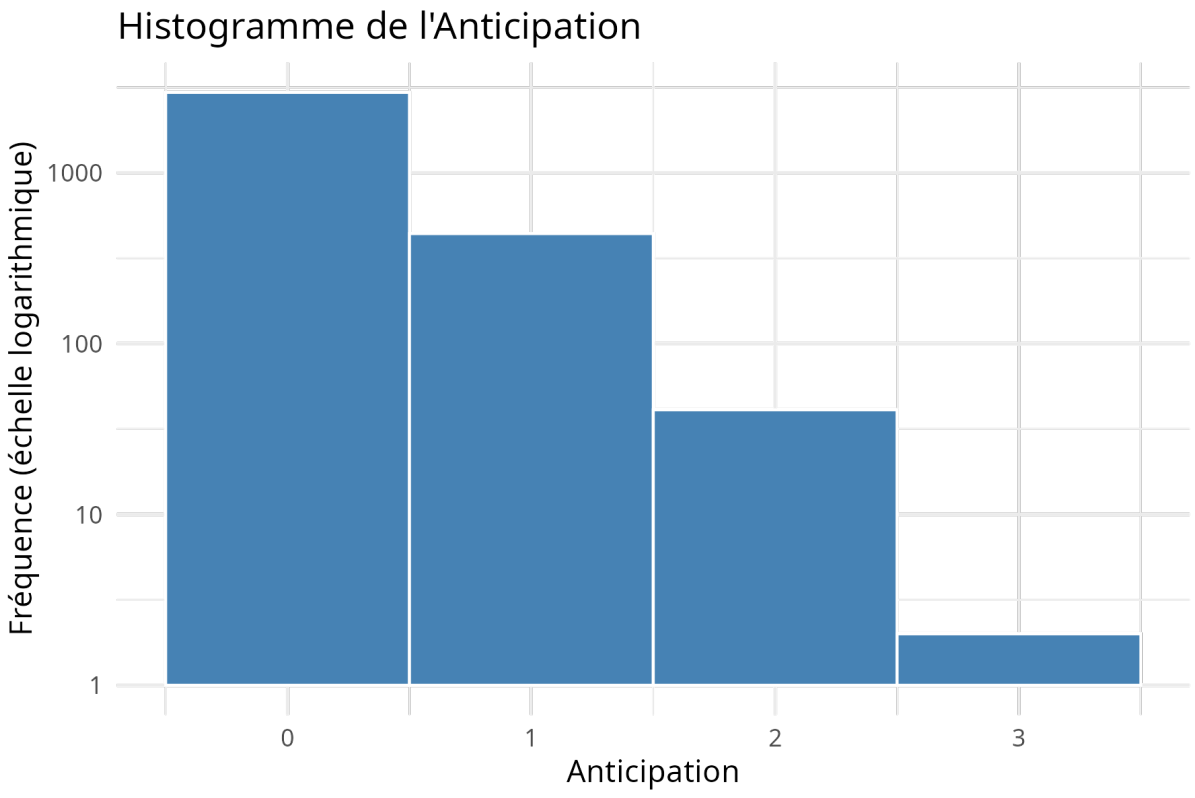

Figure 1: Histogramme de la variable **anticipation**.

2.3 Distribution des variables explicatives

La Figure 2 présente des histogrammes pour chacune des variables explicatives (émotions). L'échelle logarithmique sur l'axe des ordonnées facilite la visualisation des queues de distribution et des valeurs rares.

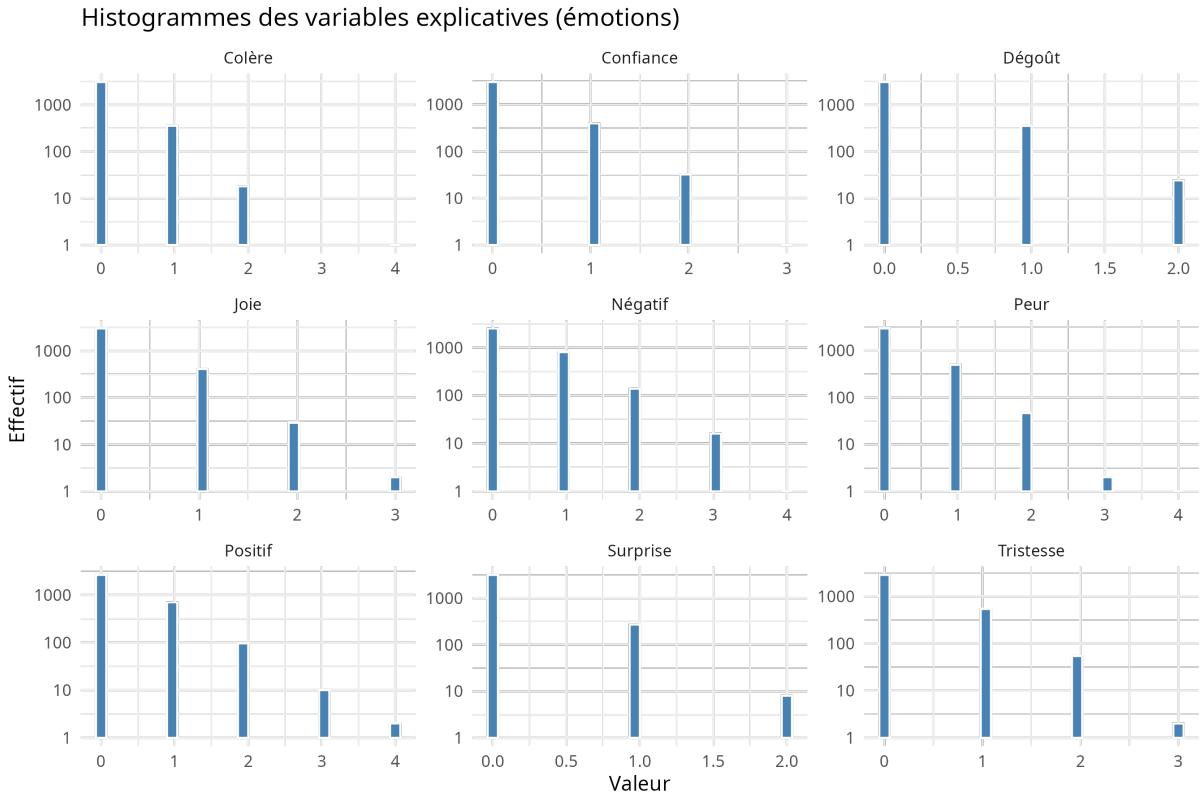

Figure 2: Histogrammes des variables explicatives (émotions), tracés en échelle logarithmique pour une meilleure lisibilité.

2.4 Corrélations avec les autres variables émotionnelles

La Figure 3 montre la matrice de corrélation entre **colère**, **anticipation**, **dégoût**, **peur**, **joie**, **tristesse**, **surprise**, **confiance**, **négatif** et **positif**. Ces corrélations ont été calculées selon la méthode de Kendall, puis réordonnées par clustering hiérarchique pour mettre en évidence d'éventuels regroupements de variables.

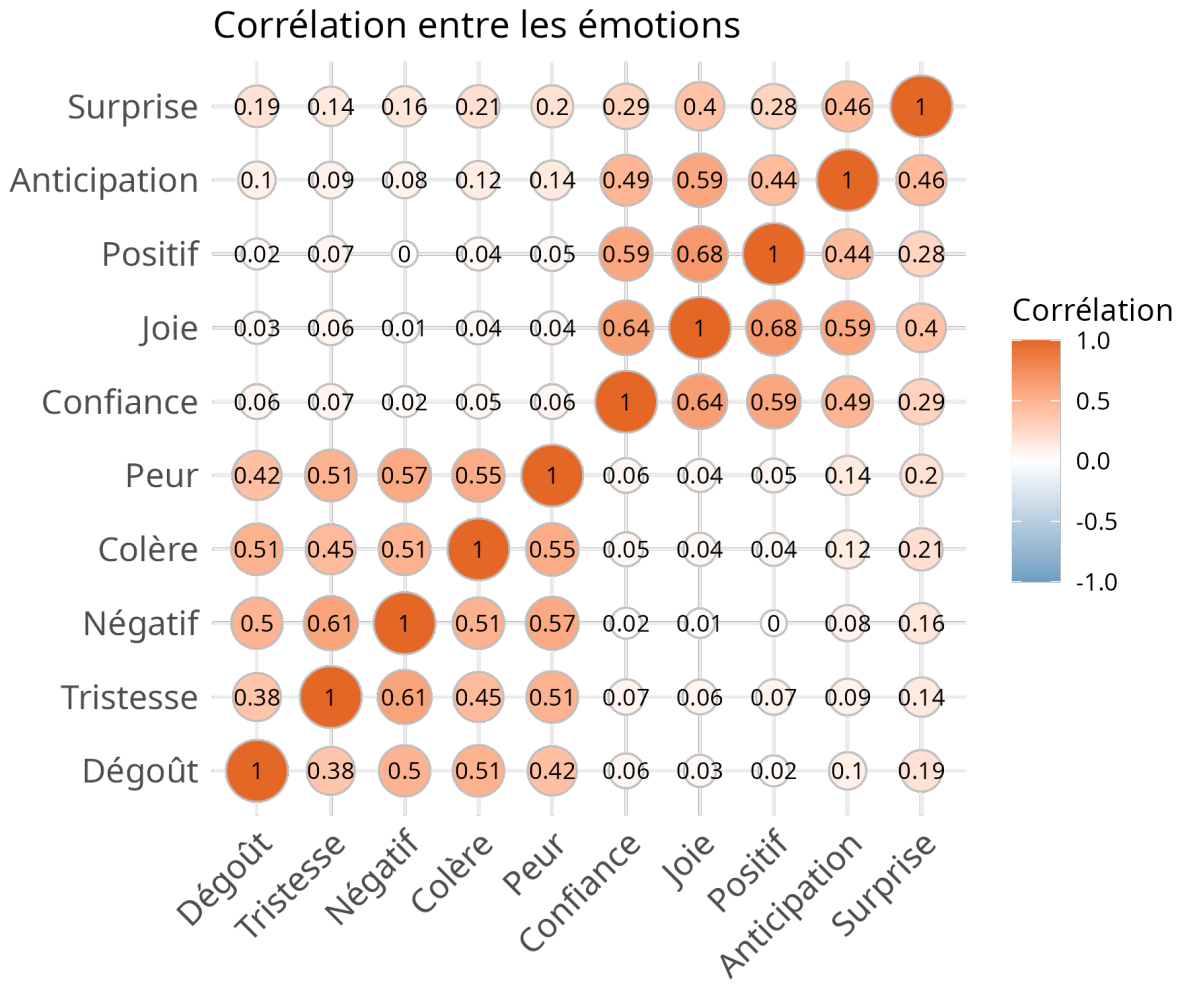

Figure 3: Matrice de corrélation (méthode : Kendall).

2.5 Fréquences d'anticipation = 0, 1, 2, 3 selon les variables catégorielles

Les tableaux ci-dessous présentent la distribution des valeurs d'**anticipation** (0, 1, 2, 3) en fonction de différentes variables catégorielles (**sexe**, **période**, **suicidal**, **heterosexual**).

Table 2: Nombre de vers ayant Anticipation = 0,1,2,3 selon le Sexe.

Sexe	Valeur_0	Valeur_1	Valeur_2	Valeur_3
Female	1516	218	21	1
Male	1443	223	20	1

Table 3: Nombre de vers ayant Anticipation = 0,1,2,3 selon la Période.

Période	Valeur_0	Valeur_1	Valeur_2	Valeur_3
Early	954	130	12	1
Later	805	150	13	1
Middle	1200	161	16	0

Table 4: Nombre de vers ayant Anticipation = 0,1,2,3 selon l'indicateur Suicidaire.

Suicidaire	Valeur_0	Valeur_1	Valeur_2	Valeur_3
FALSE	1618	194	11	0
TRUE	1341	247	30	2

Table 5: Nombre de vers ayant Anticipation = 0,1,2,3 selon l'orientation sexuelle.

heterosexual	Valeur_0	Valeur_1	Valeur_2	Valeur_3
FALSE	871	136	16	0
TRUE	1525	240	21	2
NA	563	65	4	0

3 Analyse avec réduction de la dimension

3.1 Justification du choix et prétraitement

Pour représenter plus clairement la structure multidimensionnelle des variables explicatives émotionnelles et faciliter leur interprétation, nous avons retenu l'*analyse en composantes principales* (ACP). Celle-ci présente plusieurs intérêts :

- **Réduction de la dimension :** En présence de nombreuses variables potentiellement corrélées (par exemple **colère**, **peur**, **dégoût**, etc.), l'ACP permet de synthétiser l'information sur quelques axes principaux tout en conservant l'essentiel de la variance.

- **Visualisation :** Les deux ou trois premières composantes principales offrent une représentation plus accessible, permettant de repérer des tendances ou groupements dans l'espace des données.

- **Facilité d'interprétation :** Les poids (ou *loadings*) associés aux composantes principales mettent en évidence les variables qui contribuent le plus à la variabilité globale, révélant souvent des oppositions claires (émotions négatives vs. émotions positives, etc.).

Avant de procéder à l'ACP, **toutes les variables numériques ont été centrées et réduites** de manière à leur donner une échelle comparable. Ce prétraitement évite qu'une variable à variance élevée ne domine artificiellement les composantes.

3.2 ACP et éboulis des valeurs propres

La Figure 4 présente l'éboulis (*scree plot*) des valeurs propres de l'ACP, qui illustre la part de variance expliquée par chaque axe principal. On observe que les deux premiers axes expliquent environ **63,0%** de la variance totale (**36,1 %** pour l'axe 1 et **26,9 %** pour l'axe 2), ce qui justifie l'utilisation d'un biplot bidimensionnel pour interpréter la majorité de l'information.

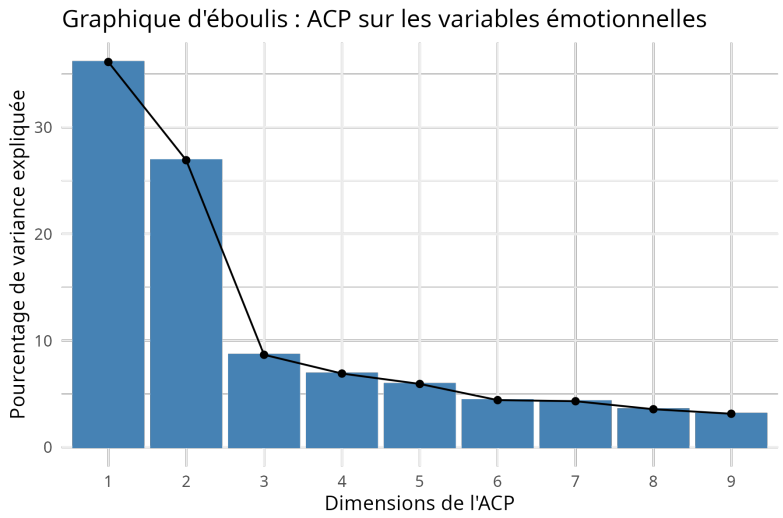

Figure 4: Éboulis (*scree plot*) des composantes principales

3.3 Biplot de l'ACP et interprétation

La Figure 5 montre le *biplot* des deux premières composantes, sur lequel les observations (chaque vers) sont projetées en points et les variables explicatives (émotions) en flèches. Pour faciliter la lecture, les points sont colorés en fonction de la valeur d'**anticipation**.

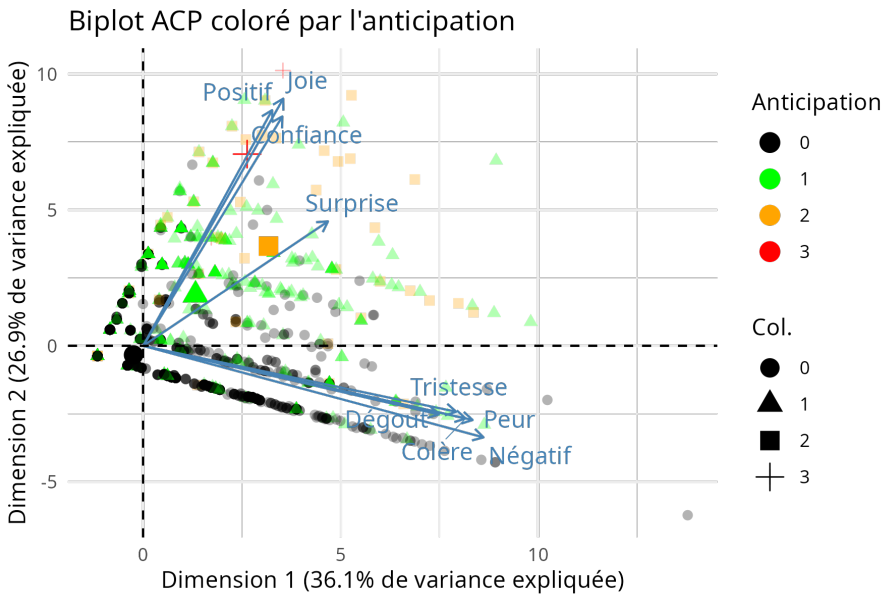

Figure 5: *Biplot* de l'ACP, coloré selon la variable **anticipation**

Interprétation des axes.

- **Axe 1 :** Il représente la croissance de l'expression des émotions, peu importe leur nature.
- **Axe 2 :** Il oppose les émotions positives aux émotions négatives.

Position de la variable anticipation. Sur le *biplot*, **anticipation** grandit dans la direction positive de l'axe associé aux émotions positives.

3.4 Conclusion et perspectives

- L'ACP met en évidence **deux grands pôles émotionnels**, opposant colère/peur/négatif à joie/confiance/positif, avec un **rôle intermédiaire** pour **anticipation**.
- Les deux premiers axes expliquent la majeure partie de la variance (environ **63,0 %**), ce qui facilite l'interprétation graphique et justifie l'utilisation d'un biplot pour la visualisation.
- D'après cette analyse, la **valence émotionnelle** (positif vs. négatif) représente la dimension la plus marquée dans nos données.

Pour aller plus loin, un modèle prédictif approfondi pourrait utiliser les composantes principales comme variables d'entrée afin de réduire le risque de *surapprentissage* et faciliter l'interprétation. D'autres méthodes de réduction de dimension, telles que `t-SNE` ou `MDS`, mériteraient également d'être explorées pour repérer d'éventuelles structures non

linéaires. Ainsi, la prochaine étape pourrait consister à comparer ces différentes approches avant de construire un modèle plus élaboré pour expliquer ou prédire **anticipation**.