Université de Tunis
École Supérieure des Sciences Économiques et Commerciales de Tunis (ESSECT)

Rapport de Projet

Analyse et Visualisation des Données de l'Enquête Nationale sur l'Emploi (T4 2023)

Réalisé par :

Turki Med Aziz Chattouti Arwa

Encadré par :

Mme **Lamia Enneifar**
Mme **Emna Souidi**

Année universitaire 2025 – 2026

Table des matières

1 Introduction et Présentation du Projet **2**
1.1 Contexte de l'Étude et Source des Données 2
1.2 Objectifs du Projet 2

2 Partie 1 : Exploration des Données et Qualité de la Base **2**
2.1 Dimensions et Structure du Jeu de Données 2
2.2 Analyse des Valeurs Manquantes 3
2.3 Identification des Valeurs Redondantes 3

3 Partie 2 : Préparation et Analyse des Chômeurs Volontaires **3**
3.1 Nettoyage et Standardisation de la Variable `Desire_to_Work` 3
3.1.1 Analyse des Modalités Existantes 3
3.1.2 Transformation des Libellés 4
3.1.3 Gestion des Valeurs Manquantes 4
3.2 Extraction du Sous-Ensemble des Chômeurs Volontaires 4
3.3 Analyse Descriptive des Chômeurs Volontaires 5
3.3.1 Répartition selon le Statut Civil 5
3.3.2 Répartition selon l'Environnement de Résidence 5

4 Partie 3 : Segmentation des Données : K-Prototypes **6**
4.1 Préparation du Dataset pour le Clustering 6
4.1.1 Nettoyage des Variables et Prétraitement 6
4.2 Méthodologie de Segmentation : K-Prototypes 7
4.2.1 Détermination du Nombre Optimal de Clusters 7
4.3 Résultats de la Segmentation 8
4.3.1 Visualisation des Clusters 8
4.3.2 Analyse des Prototypes des Clusters 9
4.3.3 Interprétation des Clusters 9

5 Conclusion et Perspectives **10**
5.1 Synthèse des Résultats Clés 10
5.2 Limites de l'Étude et Perspectives 10

1 Introduction et Présentation du Projet

1.1 Contexte de l'Étude et Source des Données

Le présent rapport documente les étapes d'un mini-projet d'analyse et de segmentation de données, basé sur l'Enquête Nationale sur l'Emploi (ENE) menée en Tunisie par l'Institut National de la Statistique (INS) pour le **quatrième trimestre de 2023**. Cette base de données brute, contenue dans le fichier `Data_T4_2023.csv`, comprend 67 444 observations pour 41 variables, offrant un aperçu détaillé des caractéristiques socio-économiques et de l'état d'activité de la population tunisienne.

1.2 Objectifs du Projet

Ce projet vise à maîtriser la chaîne de valeur d'un projet de Data Science, de la donnée brute à la conclusion interprétative. Les objectifs spécifiques sont :

1. Réaliser une **exploration descriptive** (dimension, types, valeurs manquantes) et un nettoyage minimal de la base de données.
2. **Préparer** la variable `Desire_to_Work` et isoler le sous-ensemble des individus considérés comme **chômeurs volontaires** dans la tranche d'âge [25-29] ans.
3. Effectuer une **analyse descriptive** ciblée sur ce groupe pour en identifier les caractéristiques clés (`Civil_Status`, `Environment`).
4. Mener une **segmentation non supervisée** des données en utilisant l'algorithme **K-Prototypes**, adapté aux données hétérogènes (numériques et catégorielles), et interpréter les clusters obtenus.

2 Partie 1 : Exploration des Données et Qualité de la Base

2.1 Dimensions et Structure du Jeu de Données

Le jeu de données `Data_T4_2023.csv`, issu de l'Enquête Nationale sur l'Emploi (quatrième trimestre 2023), a été chargé et exploré à l'aide du langage Python et de la bibliothèque `pandas`.

Le tableau de données contient :

— **42 668 lignes**, correspondant aux individus enquêtés ;

— **41 colonnes**, représentant les variables socio-démographiques, éducatives et professionnelles.

Les variables se répartissent en :

— **Variables numériques** : âge (`Age`), taille du ménage (`hh_size`), poids d'échantillonnage (`Weight`), année d'études (`Edu_Year`), etc.

— **Variables catégorielles** : sexe (`Gender`), région (`Region`), niveau d'éducation (`Educa_Level`), statut d'activité (`Status_Work`), secteur d'activité (`Sector_of_Activity`), entre autres.

L'inspection des premières et dernières lignes du jeu de données confirme une structure cohérente, avec des observations couvrant différentes régions du pays et des profils socio-économiques variés.

2.2 Analyse des Valeurs Manquantes

Une analyse des valeurs manquantes a été réalisée à l'aide de la fonction `df.isnull().sum()`, permettant d'identifier les colonnes présentant des données absentes.

Les résultats montrent la présence d'un nombre important de valeurs manquantes, principalement concentrées sur les variables liées à :

— l'éducation (cycle, année, diplôme) ;

— l'activité professionnelle et le statut sur le marché du travail ;

— les secteurs d'activité et les caractéristiques de l'emploi.

Le tableau 1 présente un extrait des colonnes les plus impactées par les valeurs manquantes.

TABLE 1 – Principales colonnes présentant des valeurs manquantes

Colonne	**Nombre de valeurs manquantes**	**Pourcentage (%)**
`Year_Of_Diploma`	36 037	84.45
`Sale_or_Self_cons`	42 575	99.79
`Agric_Work`	42 558	99.75
`Procedure_Look`	40 435	94.77
`Ready_Available`	39 812	93.29

Conclusion : La présence élevée de valeurs manquantes s'explique par la nature conditionnelle de plusieurs questions de l'enquête (par exemple, certaines variables ne concernent que les individus actifs ou en recherche d'emploi). Ces valeurs manquantes ont été conservées à ce stade de l'analyse, car elles portent une information implicite importante pour l'interprétation des résultats.

2.3 Identification des Valeurs Redondantes

La détection des lignes complètement dupliquées a été effectuée à l'aide de la fonction `df.duplicated().sum()`.

Les résultats indiquent :

— **0 ligne redondante** dans le jeu de données.

Ainsi, aucune suppression n'a été nécessaire à ce niveau, et l'intégrité des observations a été préservée pour les analyses ultérieures.

3 Partie 2 : Préparation et Analyse des Chômeurs Volontaires

3.1 Nettoyage et Standardisation de la Variable `Desire_to_Work`

3.1.1 Analyse des Modalités Existantes

L'exploration initiale de la variable `Desire_to_Work` a permis d'identifier trois modalités distinctes :

— `1.Yes`

— `2.No`

— des valeurs manquantes (`NaN`)

L'analyse des fréquences montre que la majorité des observations correspondent soit à une non-réponse, soit à un refus explicite de travailler.

3.1.2 Transformation des Libellés

Afin d'améliorer la lisibilité et l'exploitation de la variable, les modalités ont été standardisées comme suit :

— `1.Yes` → `Yes`

— `2.No` → `No`

Cette transformation permet une interprétation plus intuitive des réponses.

3.1.3 Gestion des Valeurs Manquantes

La colonne `Desire_to_Work` contenait initialement **24 798 valeurs manquantes**. Ces valeurs ont été remplacées par la modalité explicite `Not specified`.

Ce choix méthodologique permet :

— de conserver l'ensemble des observations ;

— de distinguer clairement une absence de réponse d'un refus explicite de travailler ;

— d'éviter une perte d'information liée à la suppression des lignes concernées.

Après traitement, la distribution finale de la variable est la suivante :

— `Not specified` : 24 798 individus

— `No` : 17 194 individus

— `Yes` : 676 individus

3.2 Extraction du Sous-Ensemble des Chômeurs Volontaires

Conformément à l'énoncé, les **chômeurs volontaires** ont été définis comme les individus :

— âgés entre **25 et 29 ans inclus** ;

— déclarant explicitement ne pas souhaiter travailler (`Desire_to_Work = No`).

Le filtre appliqué peut être formalisé comme suit :

$$(\texttt{Age} \in [25, 29]) \ \wedge \ (\texttt{Desire\_to\_Work} = \texttt{No})$$

L'application de ce filtre a permis d'extraire un sous-ensemble de **529 individus**. Le fichier `chomeurs_volontaires.csv` a été généré et contient les variables suivantes :

— `Age`

— `Civil_Status`

— `Environment`

— `Desire_to_Work`

— `Region`

— `Gender`

— `Educa_Level`

3.3 Analyse Descriptive des Chômeurs Volontaires

3.3.1 Répartition selon le Statut Civil

La répartition des chômeurs volontaires selon le statut civil est présentée dans le tableau 2.

TABLE 2 – Répartition du Statut Civil des chômeurs volontaires (25–29 ans)

Statut Civil	Effectif	Fréquence (%)
Marié(e)	281	53.1
Célibataire	240	45.4
Divorcé(e)	6	1.1
Veuf(ve)	2	0.4

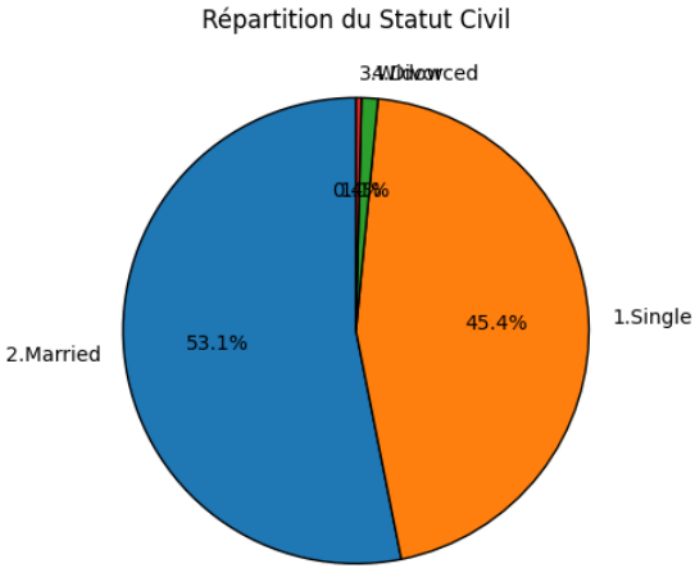

FIGURE 1 – Répartition du statut civil chez les chômeurs volontaires âgés de 25 à 29 ans.

Interprétation : Les individus mariés représentent une légère majorité des chômeurs volontaires (53.1 %), suivis de près par les célibataires (45.4 %). Cette répartition suggère que le refus de travailler dans cette tranche d'âge n'est pas exclusivement lié au statut matrimonial, mais peut résulter d'autres facteurs sociaux ou économiques.

3.3.2 Répartition selon l'Environnement de Résidence

La distribution des chômeurs volontaires selon l'environnement de résidence est illustrée dans la figure 2.

— Milieu urbain : 311 individus (58.8 %)

— Milieu rural : 218 individus (41.2 %)

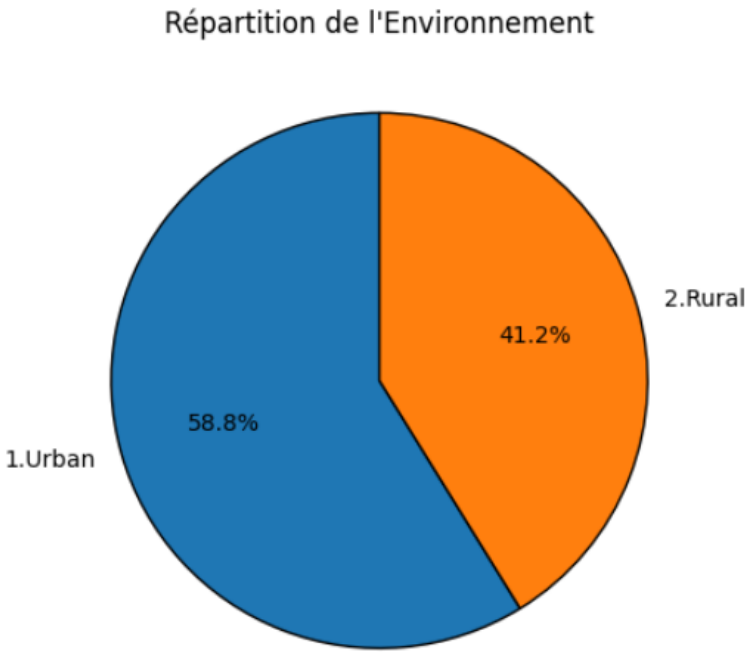

FIGURE 2 – Répartition des chômeurs volontaires selon l'environnement (urbain/rural).

Interprétation : La majorité des chômeurs volontaires réside en milieu urbain, ce qui peut s'expliquer par une concentration plus élevée des opportunités d'emploi, mais également par des attentes professionnelles plus sélectives ou un décalage entre l'offre et la demande sur le marché du travail urbain.

4 Partie 3 : Segmentation des Données : K-Prototypes

4.1 Préparation du Dataset pour le Clustering

La segmentation a été réalisée à partir du fichier `chomeurs_volontaires.csv` obtenu précédemment. Un sous-échantillon de **150 individus** a été extrait afin de faciliter l'analyse et réduire la complexité computationnelle.

Les variables initialement retenues sont :

- `Region`
- `Gender`
- `Age`
- `Educa_Level`
- `Civil_Status`

4.1.1 Nettoyage des Variables et Prétraitement

Les étapes de préparation suivantes ont été appliquées :

1. **Suppression de la variable** `Region` : L'analyse descriptive a montré que la variable `Region` présentait une très faible variabilité dans l'échantillon extrait, avec une forte

dominance de la région *Grand Tunis*. Elle n'apportait donc pas de pouvoir discriminant suffisant pour la segmentation et a été supprimée.

2. **Suppression d'instances spécifiques** : La condition (`Age` < 8) $\wedge$ (`Educa_Level` = NaN) a été vérifiée. Aucun individu ne satisfaisait cette condition, donc **aucune ligne n'a été supprimée**.
3. **Traitement des valeurs manquantes** : Les **18 valeurs manquantes** de la variable `Educa_Level` ont été remplacées par la modalité `Not specified`, afin de conserver l'ensemble des observations tout en explicitant l'absence d'information.

À l'issue de ce prétraitement, le dataset final utilisé pour le clustering contient :

— **150 observations**

— **4 variables** :

  — `Age` (numérique)

  — `Gender` (catégorielle)

  — `Educa_Level` (catégorielle)

  — `Civil_Status` (catégorielle)

4.2 Méthodologie de Segmentation : K-Prototypes

Les méthodes de clustering classiques comme le K-Means sont limitées aux variables numériques. Or, le jeu de données étudié contient à la fois des variables numériques et catégorielles.

Pour cette raison, l'algorithme **K-Prototypes** a été retenu. Cet algorithme combine :

— le principe du **K-Means** pour les variables numériques (minimisation de la variance intra-cluster) ;

— le principe du **K-Modes** pour les variables catégorielles (minimisation des dissimilarités).

Il est donc particulièrement adapté à la segmentation de profils socio-démographiques hétérogènes.

4.2.1 Détermination du Nombre Optimal de Clusters

La méthode du **Coude (Elbow Method)** a été appliquée en faisant varier le nombre de clusters K de 1 à 10. Pour chaque valeur de K, le coût intra-cluster a été calculé.

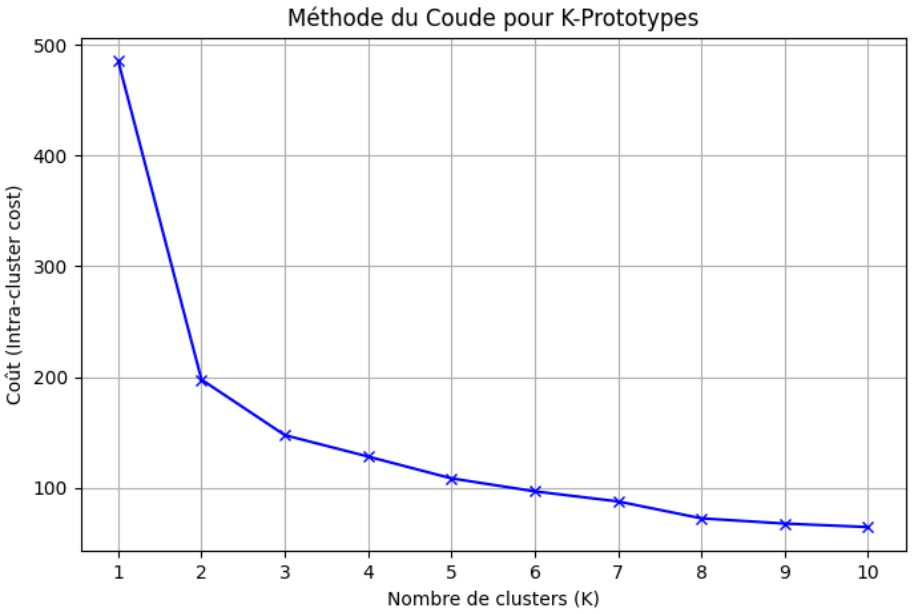

FIGURE 3 – Méthode du Coude pour la détermination du nombre optimal de clusters.

Justification du choix : L'observation du graphe montre une diminution rapide du coût jusqu'à $K = 3$, suivie d'un ralentissement marqué. Ainsi, le nombre optimal de clusters retenu est :

$$K_{\text{best}} = 3$$

4.3 Résultats de la Segmentation

L'algorithme K-Prototypes a été exécuté avec $K = 3$. Chaque individu a été affecté à un cluster en fonction de la similarité de son profil socio-démographique.

4.3.1 Visualisation des Clusters

La figure 4 illustre la répartition des individus selon leur âge et leur niveau éducatif, colorés par cluster. Les symboles `X` représentent les prototypes (centroïdes pour l'âge et modes pour les variables catégorielles).

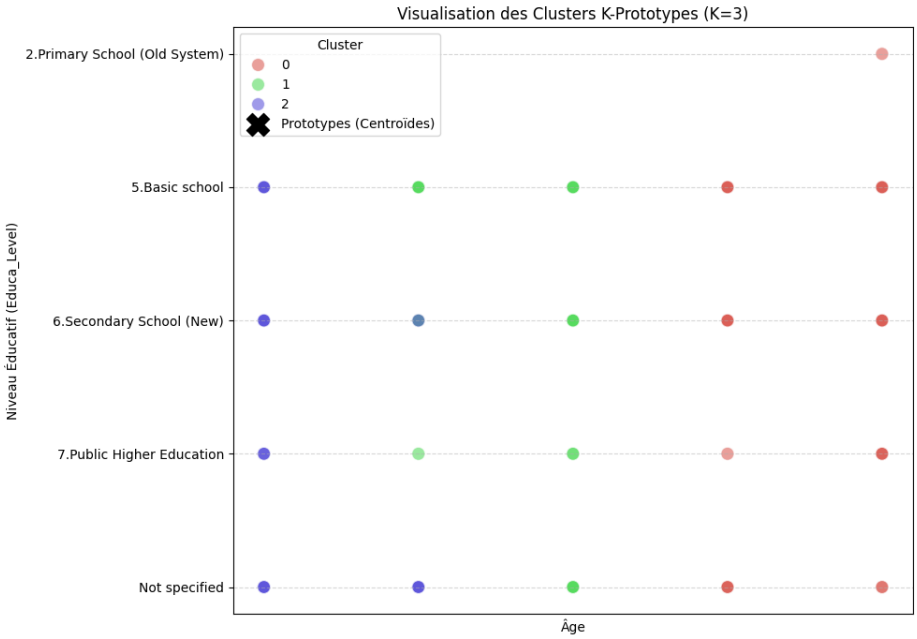

FIGURE 4 – Visualisation des clusters : Âge et Niveau Éducatif.

4.3.2 Analyse des Prototypes des Clusters

Les prototypes permettent de résumer chaque cluster par un profil type, défini par la moyenne d'âge et les modalités dominantes des variables catégorielles.

TABLE 3 – Profils types des clusters issus du K-Prototypes

Caractéristique	**Cluster 0**	**Cluster 1**	**Cluster 2**
Âge moyen	$\approx$ 26 ans	$\approx$ 27 ans	$\approx$ 28 ans
Genre dominant	Féminin	Féminin	Féminin
Niveau éducatif dominant	Not specified / Basique	Secondaire	Supérieur
Statut civil dominant	Célibataire	Marié(e)	Marié(e)

4.3.3 Interprétation des Clusters

Cluster 0 : Jeunes faiblement qualifiés ou sans information éducative

— **Profil :** Individus jeunes, majoritairement célibataires, avec un niveau éducatif bas ou non précisé.

— **Hypothèse d'inactivité :** Découragement face au marché du travail ou absence de qualifications suffisantes.

Cluster 1 : Adultes jeunes mariés de niveau secondaire

— **Profil :** Individus légèrement plus âgés, majoritairement mariés, disposant d'un niveau secondaire.

— **Hypothèse d'inactivité :** Contraintes familiales ou attentes salariales incompatibles avec l'offre disponible.

Cluster 2 : Jeunes diplômés de l'enseignement supérieur

— **Profil :** Individus proches de 29 ans, souvent mariés, avec un niveau d'éducation supérieur.
— **Hypothèse d'inactivité :** Recherche d'opportunités correspondant au niveau de qualification ou inadéquation entre formation et marché du travail.

5 Conclusion et Perspectives

5.1 Synthèse des Résultats Clés

Cette étude avait pour objectif d'analyser et de segmenter les profils des chômeurs volontaires âgés de 25 à 29 ans à partir des données de l'Enquête Nationale sur l'Emploi.

— L'analyse descriptive (Partie 2) a mis en évidence que les chômeurs volontaires de la tranche d'âge **[25–29] ans** sont majoritairement des **femmes**, avec un âge moyen d'environ **26,8 ans**. La majorité d'entre eux possède un **niveau éducatif basique ou secondaire**, bien qu'une proportion non négligeable ait atteint l'enseignement supérieur.
— La segmentation par l'algorithme **K-Prototypes** a permis d'identifier **trois profils distincts** de chômeurs volontaires. Ces clusters diffèrent principalement selon l'âge moyen, le niveau éducatif et le statut civil.
— Le cluster regroupant les individus les plus diplômés apparaît comme particulièrement intéressant du point de vue des **politiques d'emploi**, car il met en évidence un potentiel de compétences non exploité, suggérant un décalage entre l'offre de travail qualifiée et les opportunités disponibles sur le marché.

5.2 Limites de l'Étude et Perspectives

— **Limites :**
  — La segmentation K-Prototypes a été réalisée sur un **sous-échantillon limité de 150 observations**, ce qui peut restreindre la généralisation des résultats à l'ensemble de la population étudiée.
  — Le remplacement des valeurs manquantes de la variable `Educa_Level` par la modalité *Not specified* peut introduire un **biais d'interprétation** dans la caractérisation des clusters.
  — Le choix du nombre optimal de clusters K_{best} repose sur l'interprétation visuelle de la méthode du coude, ce qui introduit une part de **subjectivité**.
— **Perspectives :**
  — Étendre l'analyse à un **échantillon plus large** ou à l'ensemble des données disponibles afin d'améliorer la robustesse des résultats.
  — Intégrer d'autres variables explicatives, telles que `Reason_Not_Looking` ou l'expérience professionnelle, si les données sont suffisamment complètes.
  — Comparer les profils obtenus avec ceux d'autres périodes (par exemple **T1 2024**) afin d'analyser l'évolution du chômage volontaire dans le temps.
  — Tester d'autres techniques de segmentation, comme **DBSCAN** ou le **clustering hiérarchique**, pour comparer la stabilité et la pertinence des clusters obtenus.