Cahier des charges du mini-projet d'analyse statistique avec R

Abderrahmen Youssef — Groupe DS4

1. Titre du projet

Comprendre la production de diplômés en Tunisie : analyse par domaines, spécialités et établissements.

2. Problématique

Comment se répartissent les diplômés de l'enseignement supérieur public en Tunisie selon les domaines d'étude, les établissements, les universités et le genre, et quelles structures ou déséquilibres peut-on mettre en évidence à partir des données publiques disponibles ?

3. Objectifs

- Construire un jeu de données propre et lisible à partir de la table `fact_diplomes` et des tables de correspondance des codes (universités, établissements, domaines, diplômes).
- Étudier la répartition des diplômés par sexe en fonction des domaines, des spécialités et des établissements.
- Mettre en évidence les domaines et groupes de formation les plus « productifs » en termes de nombre de diplômés.
- Comparer les profils de formation des universités et des établissements (spécialisation, diversité de l'offre de formation, etc.).
- Analyser la structure globale du système de formation à l'aide d'outils d'analyse statistique (analyses croisées et analyse multivariée).

4. Données utilisées

Le projet repose sur l'exploitation d'un ensemble de jeux de données officiels décrivant la production de diplômés de l'enseignement supérieur public tunisien. L'architecture globale suit un modèle en étoile : une table principale contenant les mesures quantitatives, et plusieurs tables de correspondance permettant d'enrichir et d'interpréter les codes présents dans la table centrale.

1. Table principale : `fact_diplomes`

Cette table regroupe les informations quantitatives relatives au nombre de diplômés. Elle contient les variables suivantes :

- `universite_code` : code de l'université d'affectation,
- `etablissement_code` : code de l'établissement (faculté, institut, école…),
- `domain_code` : code du domaine d'étude,
- `diplome_code` : code du type de diplôme,
- `speciality` : spécialité de formation,

— `group` : groupe ou famille de spécialités,
— `diplomes_M` : nombre de diplômés hommes,
— `diplomes_F` : nombre de diplômées femmes,
— `diplomes_total` : nombre total de diplômés.

2. Tables de correspondance (dimensions)

Ces tables permettent de remplacer les codes présents dans `fact_diplomes` par leurs libellés descriptifs :

— **uni_code** : association entre le code université et le nom complet de l'université,
— **etablissment_code** : liste des établissements et leurs libellés,
— **domaines_code** : correspondance entre les codes et les libellés des domaines d'étude,
— **diplomes_code** : nomenclature des diplômes (ex. Licence fondamentale, Diplôme national d'ingénieur, Mastère professionnel, Doctorat...).

3. Structure finale du jeu de données fusionné

Après la fusion des tables de correspondance avec la table `fact_diplomes`, chaque ligne du dataset final contient :

— l'université et l'établissement d'enseignement,
— le domaine et la spécialité de formation,
— le groupe de spécialité ou famille disciplinaire,
— le type de diplôme (Licence, Mastère, Ingénieur, Doctorat...),
— le nombre de diplômés par sexe (H/F) ainsi que le total.

Aperçu du dataset final fusionné

L'extrait ci-dessous illustre la structure attendue du jeu de données.

Ligne 1
Université : Université de Tunis
Établissement : ISG Tunis
Domaine : Sciences économiques
Spécialité : Finance
Groupe : Sciences de gestion
Diplôme : Licence fondamentale
Diplômés H : 120
Diplômées F : 180
Total : 300

Ligne 2
Université : Université de Sfax
Établissement : ENIS
Domaine : Informatique
Spécialité : Génie logiciel
Groupe : Ingénierie
Diplôme : Diplôme national d'ingénieur
Diplômés H : 85
Diplômées F : 45
Total : 130