Réalisé par : Supervised by:  
 AITOUMEZOUAR Asmaa ZAHIR Jihad

BELLANI Ayoub

EL HILALI Mohamed Amine

Boujad Fatima Zahra

Tableau de Bord BI à la population marocaine

Sommaire

[Introduction 3](#_Toc187263370)

[Chapitre1 : Analyse des besoins et Collection des données 4](#_Toc187263371)

[1. Définition des besoins 5](#_Toc187263372)

[2. Identification des sources de donnees 5](#_Toc187263373)

[3. Collecte des données 7](#_Toc187263374)

[4. Nettoyage et organisation des données 9](#_Toc187263375)

[Chapitre2 : Organisation et transformation des données 10](#_Toc187263376)

[1. Introduction au Schéma de Constellation et au Processus ETL 11](#_Toc187263377)

[2. Modélisation Multidimensionnelle 11](#_Toc187263378)

[3. Extraction, Transformation, Chargement (ETL) 12](#_Toc187263379)

[3.1. Chargement des Dimensions et Sous-dimensions 12](#_Toc187263380)

[3.2. Transformation et Chargement des Données de la Table des Faits 13](#_Toc187263381)

[3.2.1. Organisation des fichiers Excel 13](#_Toc187263382)

[3.2.2. Extraction : 14](#_Toc187263383)

[3.2.3. Transformation 14](#_Toc187263384)

[3.2.4. Chargement : 14](#_Toc187263385)

[Chapitre3 : Création des visualisations interactives et automatisation 16](#_Toc187263386)

[1. Création des Tableaux de Bord 17](#_Toc187263387)

[2. Stratégie perspective de Mise à Jour Périodique 18](#_Toc187263388)

[3. Résultats Obtenus 18](#_Toc187263389)

[4. Analyse des Résultats 18](#_Toc187263390)

[Conclusion 20](#_Toc187263391)

Liste de Figure

[Figure 1:site HCP 7](#_Toc187270834)

[Figure 2: site DataGov.ma 7](#_Toc187270835)

[Figure 3: Matrice de bus des données 8](#_Toc187270836)

[Figure 4: fichier téléchargé .xls 9](#_Toc187270837)

[Figure 5 :fichier après les modifications code python 10](#_Toc187270838)

[Figure 6: Schéma de base de données 13](#_Toc187270839)

[Figure 7:transformation des tables de dimension 14](#_Toc187270840)

[Figure 8:l'emplacement des fichiers 15](#_Toc187270841)

[Figure 9: Transformation pour la liste des fichiers 16](#_Toc187270842)

[Figure 10: Transformation des données de Démographie et Handicap 17](#_Toc187270843)

[Figure 11: Transformation des données de Emploi 17](#_Toc187270844)

[Figure 12: Transformation des données de Education 17](#_Toc187270845)

[Figure 13: Transformation des données de santé 18](#_Toc187270846)

[Figure 14: Job\_Automatisation\_transformation des fichier 18](#_Toc187270847)

[Figure 15:distrubution de population géographiquement 20](#_Toc187270848)

[Figure 16:Evolution au niveau de sante 21](#_Toc187270849)

[Figure 17:Visualisation, Analyse d'Education 21](#_Toc187270850)

[Figure 18:analyse au niveau d'Emploi 22](#_Toc187270851)

[Figure 19: Analyse au niveau Famille 22](#_Toc187270852)

# Introduction

Dans le cadre de l’apprentissage des outils et méthodes de la Business Intelligence (BI), l’objectif est de construire un tableau de bord interactif pour analyser les données liées à la population marocaine. Ce tableau de bord permettra de mieux comprendre des aspects importants comme la démographie, l’éducation, l’emploi, la famille, et le handicap. Grâce à des outils modernes de BI, nous pouvons transformer des données complexes en visualisations simples et utiles pour les prises de décisions.

Le projet suit une série d’étapes claires pour atteindre cet objectif :

1. Identifier les besoins et collecter les données nécessaires.
2. Organiser et transformer les données en utilisant des outils adaptés.
3. Créer des tableaux de bord interactifs pour rendre les informations facilement compréhensibles.
4. Mettre en place un système pour mettre à jour les données régulièrement et automatiquement.

Ce rapport est structuré en plusieurs chapitres qui détaillent chaque étape du projet, des besoins initiaux à la présentation des résultats. L’objectif est non seulement de montrer comment les outils de BI ont été utilisés, mais aussi de démontrer comment ces compétences peuvent être appliquées à des projets concrets et utiles.

# Chapitre1 : Analyse des besoins et Collection des données

La première étape de tout projet de Business Intelligence (BI) consiste à comprendre ce que l’on souhaite analyser et quelles données sont nécessaires pour y parvenir. Ce chapitre explique comment nous avons identifié les besoins en données, choisi les sources les plus adaptées et collecté les informations nécessaires pour construire le tableau de bord.

## Définition des besoins

Avant de commencer à chercher des données, il est important de définir clairement les besoins du projet. Dans notre cas, nous avons décidé d’analyser plusieurs aspects de la population marocaine, comme :

**Démographie** : La taille de la population, répartie par âge, sexe, et milieu urbain ou milieu de type rural. L’analyse inclut également les changements observés d’année en année.

**Famille** : Les tendances concernant l’âge moyen au premier mariage, le taux de célibat définitif, et le nombre de divorces, répartie par âge, sexe, et milieu urbain ou milieu de type rural et année.

**Éducation** : Le taux d’analphabétisme, avec une comparaison entre hommes et femmes et entre les milieux urbains et ruraux. Cette analyse est effectuée sur plusieurs années pour observer les évolutions.

**Emploi** : L’évolution du taux d’activité et aussi du taux de chômage, avec une comparaison entre hommes et femmes et entre les milieux urbains et ruraux, et année.

**Hôpital** : L’évolution des nombres des hôpitaux, des pharmacies, des docteurs et des capacités, avec une comparaison par les années.

**Handicap** : La taille de la population en situation de handicap, répartie par âge, sexe, et milieu urbain ou milieu de type rural, et année.

Ces besoins définissent les indicateurs clés qui seront utilisés dans notre tableau de bord.

## Identification des sources de données

Pour répondre aux besoins définis, nous avons sélectionné des sources fiables et accessibles, fournissant des données pertinentes sur la population marocaine. Les principales sources utilisées sont :

* **Site du Haut-Commissariat au Plan (HCP)** : Ce site fournit une large gamme de données statistiques sur des aspects démographiques, économiques, et sociaux. Nous avons extrait des fichiers contenant des informations détaillées sur la démographie, l’éducation, l’emploi, la famille, et le handicap. Ces données sont organisées en séries chronologiques, ce qui facilite l’analyse des évolutions au fil des années.

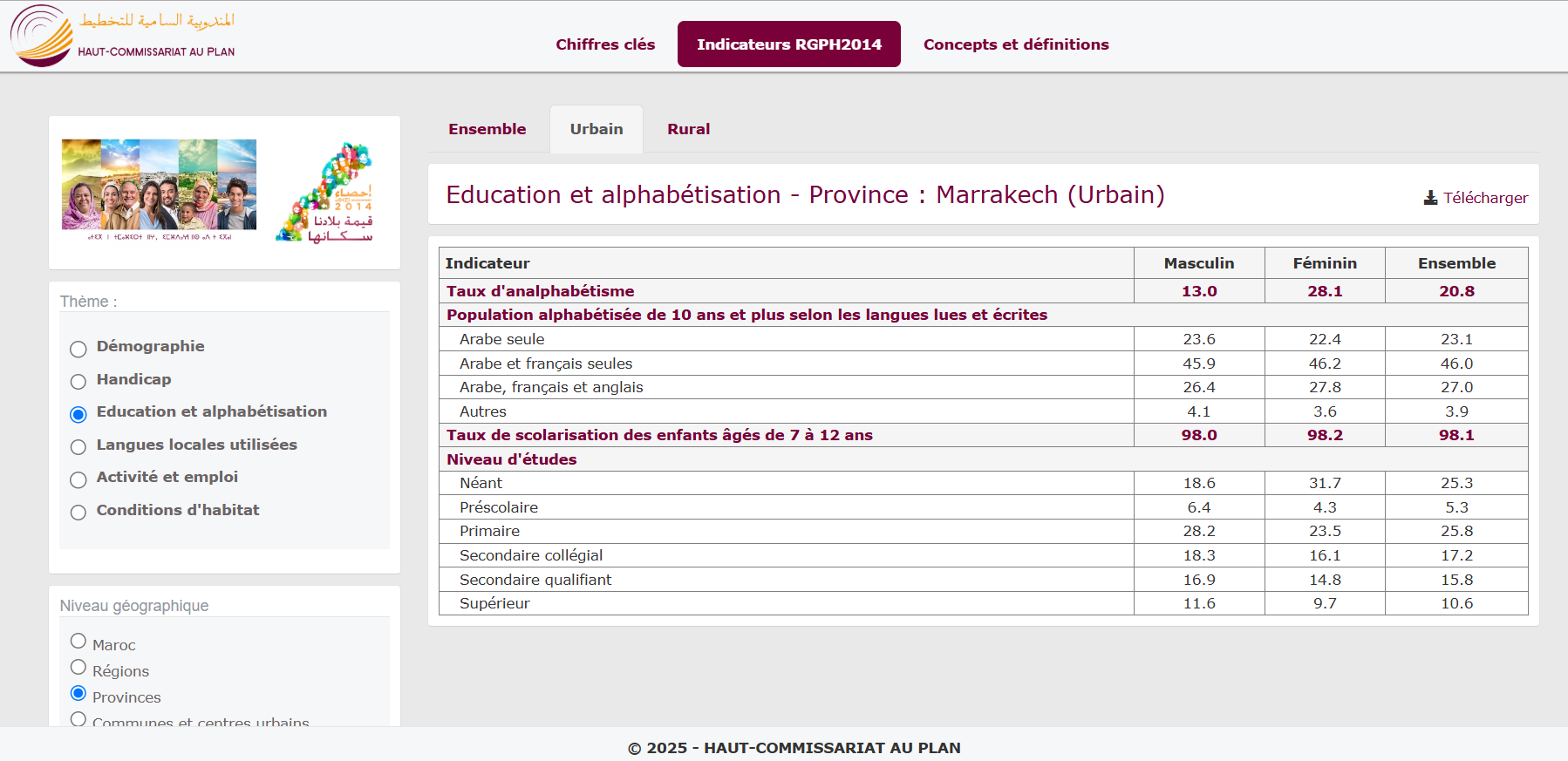


Figure 1:site HCP

* **Portail DataGov.ma** : Ce portail est une plateforme officielle qui regroupe des données ouvertes issues de divers départements gouvernementaux marocains. Nous y avons trouvé des fichiers complémentaires permettant de compléter les informations obtenues depuis le site du HCP, notamment des indicateurs spécifiques liés à l’éducation, à l’emploi, et à la population en situation de handicap.

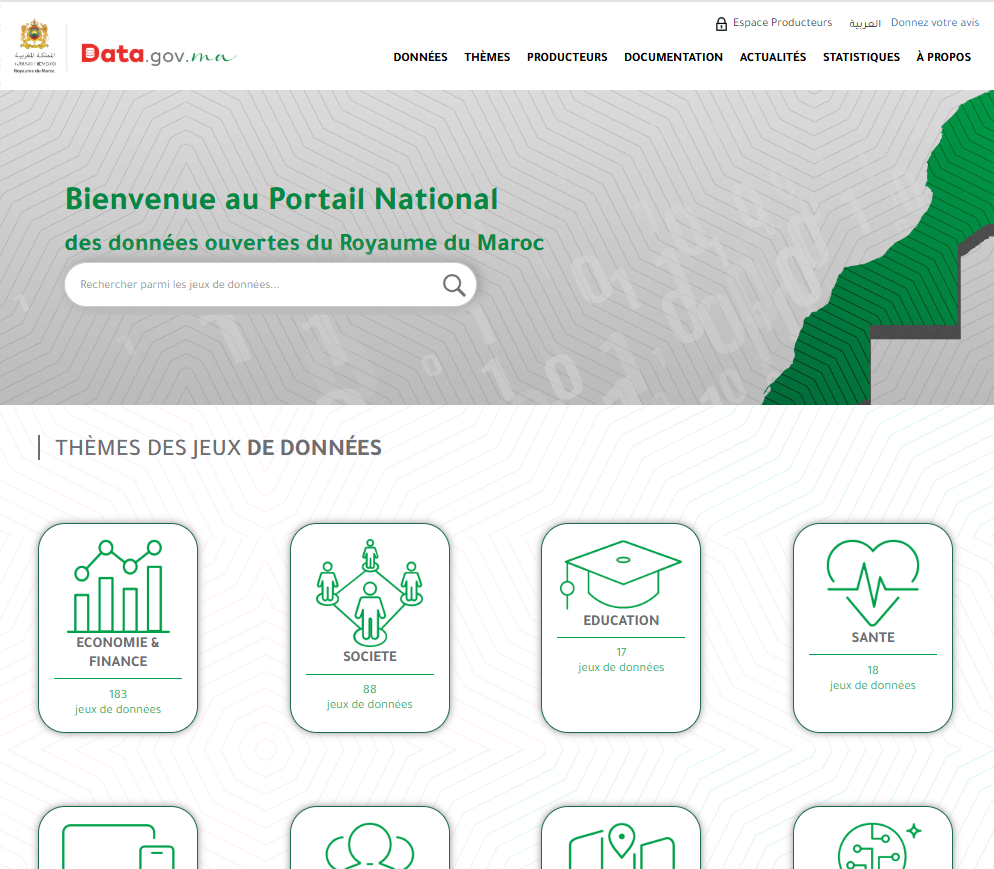


Figure 2: site DataGov.ma

Pour garantir la qualité et la pertinence des données utilisées dans ce projet, nous avons évalué les sources sélectionnées selon plusieurs critères de fiabilité. Les deux principales sources utilisées sont le site du Haut-Commissariat au Plan (HCP) et le portail DataGov.ma. La matrice suivante résume cette évaluation : Ces deux sources ont été choisies pour leur fiabilité et leur richesse en données structurées, indispensables pour mener à bien le projet. Leur utilisation garantit également une représentativité et une actualité des données analysées.

|  |  |  |
| --- | --- | --- |
| Critère | HCP (Haut-Commissariat au Plan) | DataGov.ma |
| Accessibilité | Publique, gratuite, accès direct en ligne. | Publique, gratuite, accès direct en ligne. |
| Exactitude | Données fiables, recueillies par des institutions officielles. | Données fiables, validation par des départements gouvernementaux. |
| Actualité | Mise à jour annuelle ou régulière (selon les séries). | Mises à jour régulières mais moins fréquentes que le HCP. |
| Couverture | Très large : démographie, éducation, économie, emploi, etc. | Large mais moins étendue sur certaines thématiques, surtout sur l'éducation et l'emploi. |
| Méthodologie | Méthodes statistiques rigoureuses, détails fournis. | Méthodes transparentes mais parfois moins détaillées. |
| Utilisation antérieure | Référencé dans de nombreuses études et rapports officiels. | Utilisé par des chercheurs et institutions, mais moins d’historique que le HCP. |
| Défauts | Complexité dans le format des fichiers , rendant le traitement plus difficile | Contenu parfois limité avec des fichiers ne contenant qu'une seule information |

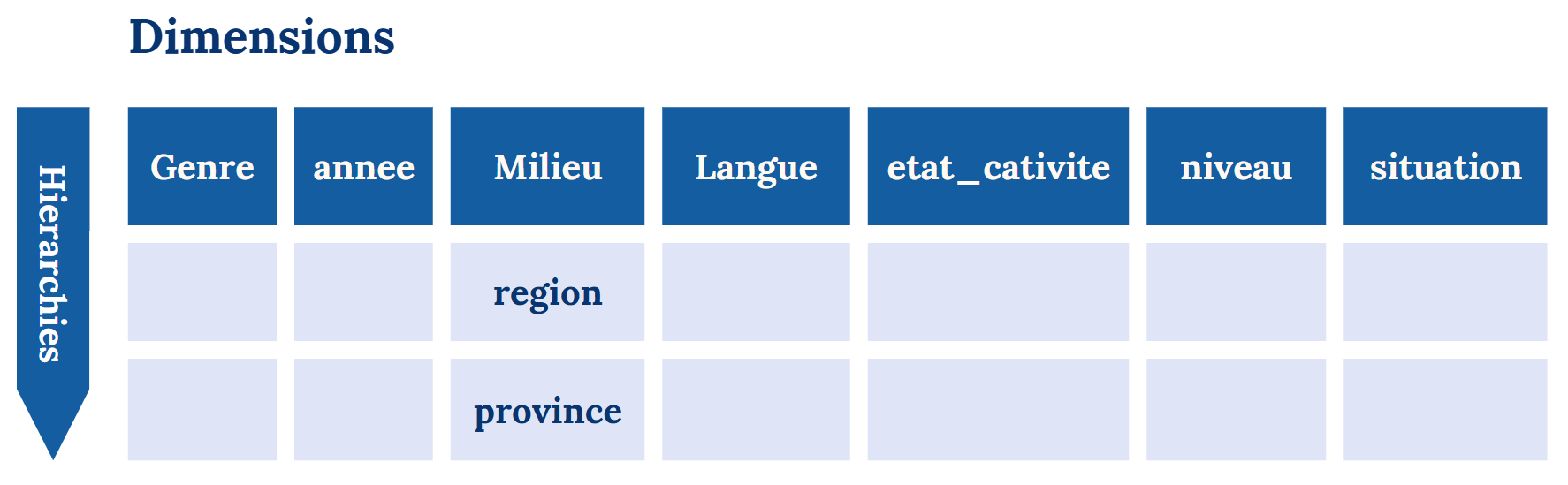


Figure 3: Matrice de bus des données

## Collecte des données

Pour collecter les données nécessaires à notre projet, nous avons utilisé des outils automatiques afin de simplifier et d’accélérer le processus. Au lieu de télécharger manuellement les fichiers depuis les sites du HCP et de DataGov.ma, nous avons développé un **code en Python** pour effectuer de **l’extraction de données automatique**. Ce script a permis de parcourir les pages web de ces sites, d’identifier les fichiers pertinents au format **.xls**, et de les télécharger automatiquement.

Une fois les fichiers collectés, nous avons constaté qu’ils étaient au format tableau HTML, intégré dans un fichier **.xls** sous Excel, ce qui n’est pas compatible avec l’outil ETL que nous avons choisi, à savoir **Pentaho Data Integration** (**PDI**). Pour résoudre ce problème, nous avons utilisé un autre script Python pour convertir automatiquement ces fichiers en format **.xlsx**. Ce processus de conversion a également permis de garantir une standardisation des fichiers et une compatibilité avec les outils de traitement utilisés.

En plus de la conversion, nous avons effectué quelques modifications automatiques sur les fichiers à l’aide de Python. Par exemple, des colonnes spécifiques ont été ajoutées pour enrichir les tables, comme des indicateurs des regroupements basés sur des années ou des catégories comme milieu. Ces changements ont permis d’adapter les fichiers aux besoins du projet tout en limitant les interventions manuelles.

Cette approche basée sur l’automatisation a non seulement réduit le temps de collecte et de préparation des données, mais a également amélioré leur qualité et leur uniformité, ce qui est essentiel pour garantir des analyses fiables.

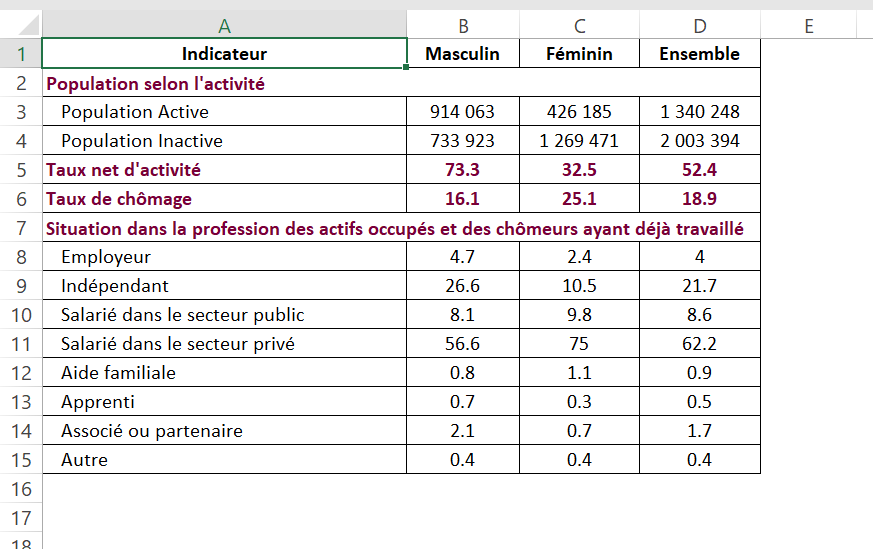


Figure 4: fichier téléchargé .xls

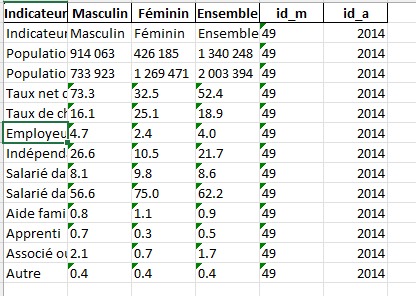


Figure 5 :fichier après les modifications code python

## Nettoyage et organisation des données

Après la collecte, les données ont été nettoyées et organisées pour garantir leur qualité et leur cohérence. Cela a impliqué la suppression des doublons et des lignes vides par exemple le tableau de données de toutes les tables de faits de province Casablanca rural est vide donc nous avons le supprimé, ainsi que la standardisation des formats, par exemple en uniformisant les noms des régions et les dates. Une attention particulière a été accordée à la vérification de la cohérence des données, notamment en s’assurant que les totaux correspondaient aux valeurs agrégées. Ce processus de nettoyage était essentiel pour éviter les erreurs lors des étapes suivantes du projet.

Ce premier chapitre a permis de poser les bases du projet en définissant les besoins en données, en identifiant des sources fiables et en collectant les informations nécessaires. Grâce à l’utilisation de Python pour automatiser une partie du processus, les données ont été organisées de manière efficace et prêtes pour les étapes suivantes, notamment la modélisation multidimensionnelle et la création des visualisations interactives. Ces efforts initiaux garantissent la qualité et la fiabilité des analyses à venir.

# Chapitre2 : Organisation et transformation des données

Une base de données bien structurée est essentielle pour organiser efficacement les données. Ce chapitre détaille la modélisation multidimensionnelle utilisée pour organiser les données dans un schéma de constellation ainsi que le processus ETL (Extraction, Transformation, Chargement) pour intégrer ces données dans une base MySQL.

## Introduction au Schéma de Constellation et au Processus ETL

Un schéma de constellation est une approche utilisée en modélisation multidimensionnelle dans les bases de données décisionnelles. Il combine plusieurs schémas en étoile dans une structure unifiée, permettant à une table de faits d’être associée à plusieurs dimensions et d’autres tables de faits. Ce schéma est particulièrement utile lorsque plusieurs analyses partagent des dimensions communes, offrant ainsi une vue intégrée et flexible des données.

Le processus ETL (Extraction, Transformation, Chargement) est une méthode essentielle pour intégrer des données dans un entrepôt de données :

* Extraction : Collecte des données à partir de différentes sources (fichiers, bases de données, etc.).
* Transformation : Nettoyage, mise en forme et adaptation des données pour qu’elles respectent le modèle cible.
* Chargement : Insertion des données transformées dans la base de données décisionnelle.

Ces concepts jouent un rôle crucial dans le développement et l’utilisation des systèmes d’aide à la décision.

## Modélisation Multidimensionnelle

La modélisation multidimensionnelle dans ce projet suit une architecture étoile, combinée avec certains aspects d’un schéma de constellation. Voici les principales dimensions et sous-dimensions identifiées :

* Dimension Année : Cette dimension contient les années associées aux données collectées.
* Dimension Genre : Décrit les catégories de genre (exemple : masculin, féminin).
* Dimension Milieu : Représente les milieux de vie (urbain, rural).
* Dimension État d’Activité : Décrit les états d’activité (actif, inactif, etc.).
* Dimension Langue : Répertorie les langues parlées.
* Dimension Âge : Contient les tranches d’âge.
* Dimension Situation : Liée au taux par situation (exemple : célibataire, marié, etc.).
* Dimension Niveau : Décrite par le taux selon le niveau d’étude (exemple : primaire, secondaire, supérieur).

Sous-dimensions Géographiques :

* Sous-dimension Province : Regroupe les provinces des données.
* Sous-dimension Région : Reliée à la sous-dimension Province et répertorie les régions.

Les relations entre ces dimensions et la table des faits permettent d’obtenir une vue multidimensionnelle des données pour une analyse efficace.

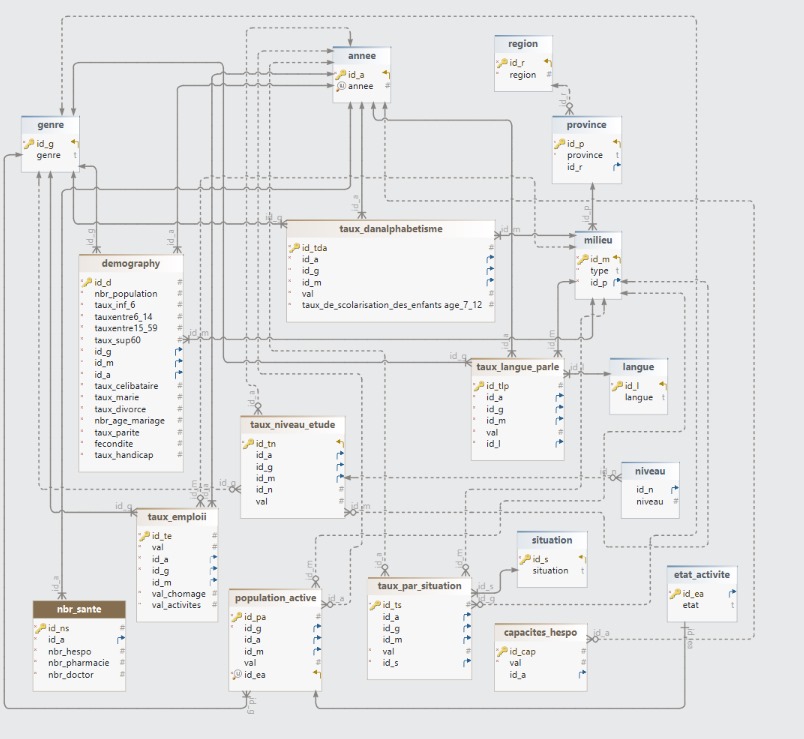


Figure 6: Schéma de base de données

## Extraction, Transformation, Chargement (ETL)

Le processus ETL est divisé en deux étapes principales :

### Chargement des Dimensions et Sous-dimensions

Pour chaque fichier qui contient les données d’un spécifique dimension ou sous\_dimension

**Extraction** :

* Les fichiers .xlsx contenant les données des dimensions et sous-dimensions par exemple, Dim\_Annee.xlsx, Dim\_Genre.xlsx, sous\_sous\_dim\_region.xlsx, et Dim\_milieu ont été importés dans PDI à l’aide du composant Excel Input.
* Chaque fichier représente une dimension ou une sous-dimension spécifique.

**Transformation** :

* Nous avons faire plusieurs transformations pour chaque fichier. Pour chaque fichier les colonnes ont été mappées aux champs correspondants dans les tables MySQL.

**Chargement** :

* Les données transformées ont été chargées dans les tables MySQL en utilisant le composant Table Output.

Exemple pour la sous-dimension region :

Le fichier region.xlsx contient les colonnes id\_r et region. Ces colonnes ont été mappées aux champs de la table region, dans la base de données.

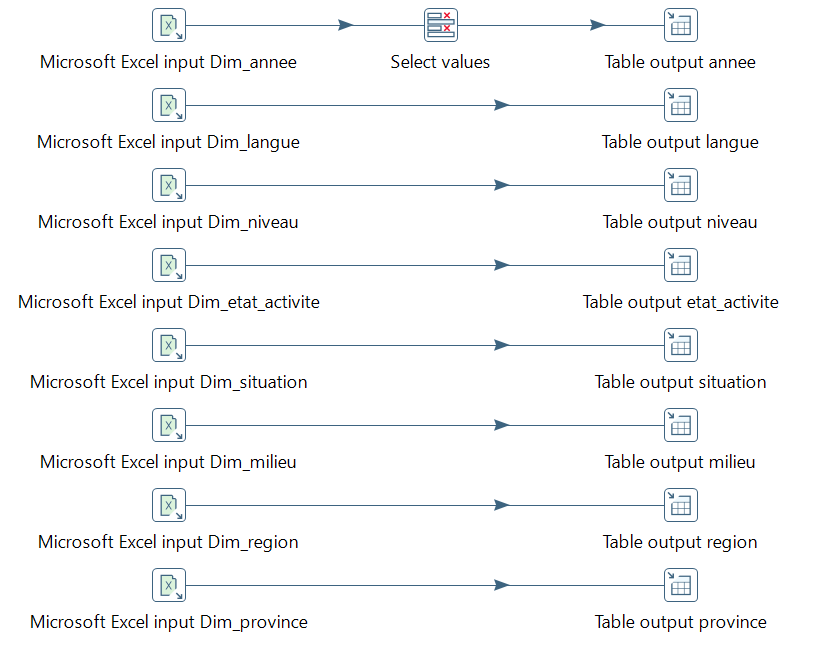


Figure 7:transformation des tables de dimension

### Transformation et Chargement des Données de la Table des Faits

#### Organisation des fichiers Excel

Les fichiers contenant les mesures pour la table des faits ont été divisés en quatre groupes selon leur thématique :

* Groupe 1 : contient les fichiers Excel contenant les données de Démographie, Famille et Handicap. Ces données correspondent à la première table de faits, appelée "demography".
* Groupe 2 : contient les fichiers Excel contenant les données d’Éducation. Ces données correspondent aux tables de faits "taux\_d'analphabetisme", "taux\_langue\_parle", et "taux\_niveau\_etude".
* Groupe 3 : contient les fichiers Excel contenant les données d’Hôpital. Ces données correspondent aux tables de faits "nbr\_sante" et "capacites\_hespo".
* Groupe 4 : contient les fichiers Excel contenant les données d’Emploi. Ces données correspondent aux tables de faits "taux\_emploii", "taux\_par\_situation", et "population\_active".

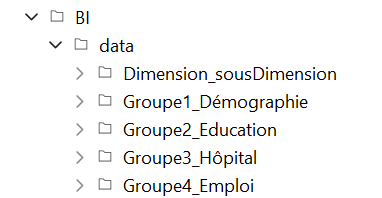


Figure 8:l'emplacement des fichiers

#### Extraction :

Les fichiers .xlsx contenant les données des tables des faits chaque ensemble des fichiers qui ont dans le même groupe est existé dans un répertoire nomme par exemple pour le répertoire qui contient les fichiers du premier groupe nomme Group1\_Démographie. Alors ces fichiers ont été importés dans PDI à l’aide du composant **Microsoft Excel Input**.

* + 1. Transformation :
* La première transformation :

Pour ne répéter pas les transformations pour chaque fichier nous avons utilisé le composant **Get File Names** qui permet de récupérer la liste des fichiers d'un dépôt Get, généralement utilisé pour identifier les fichiers à traiter dans une transformation. Nous avons aussi utilisé le composant **Copy Rows to Result** pour Copie les lignes de données de composant **Get File Names** vers le résultat, pour qu'elles soient utilisées par une transformation ou un job suivant.

* La deuxième transformation :

L'utilisation du composant **Get Rows from Result** pour récupérer les lignes de données copiées dans le résultat par la première transformation, souvent combiné avec **Copy Rows to Result**. Ce composant est suivi par le composant Set Variables, qui définit les variables correspondant aux chemins des fichiers Excel utilisés dans l’étape suivante.

Les données manquantes ou incorrectes ont été nettoyées à l’aide du composant **Filter Rows** pour éliminer les lignes invalides. Les données contiennent également des tirets "-". Nous avons lu les données comme des nombres réels, car elles avaient été lues comme des chaînes de caractères. Ensuite, nous avons supprimé les espaces et ajouté des champs calculés, en utilisant le composant **Modified Java Script Value**. Ensuite nous avons aussi utilisé le composant Select Values pour sélectionner les colonnes sur lesquelles nous voulions travailler et pour renommer les colonnes.

Le composant **Row Denormalizer** a été utilisé pour transformer des données normalisées organisées sous forme de plusieurs lignes en une structure dénormalisée organisée sous forme de colonnes. Ensuite, nous avons utilisé le composant **Row Normalizer** pour transformer une structure dénormalisée où les données sont dans des colonnes en une structure normalisée où les données sont réparties sur plusieurs lignes.

Enfin, nous avons utilisé le composant **Add Constants** pour ajouter des colonnes.

#### Chargement :

* Les données transformées ont été chargées dans les tables MySQL à l’aide du composant **Table Output**.
* Quatre jobs distincts ont été créés pour chaque groupe, commençant par la première transformation, suivie de la deuxième transformation, pour garantir une transformation et un chargement adaptés à chaque fichier.

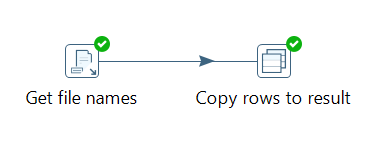


Figure 9: Transformation pour la liste des fichiers

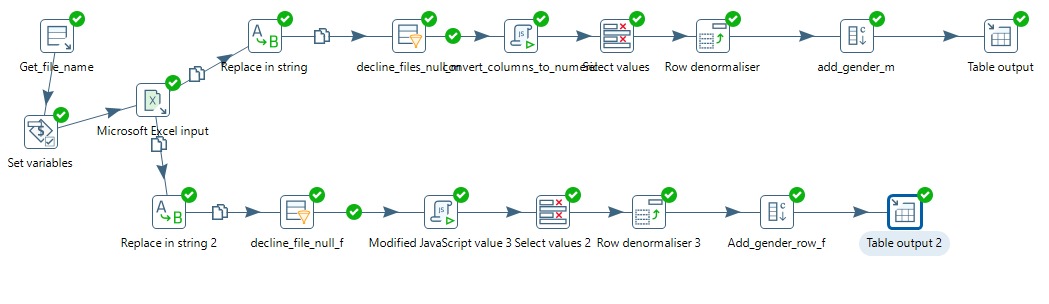


Figure 10: Transformation des données de Démographie et Handicap

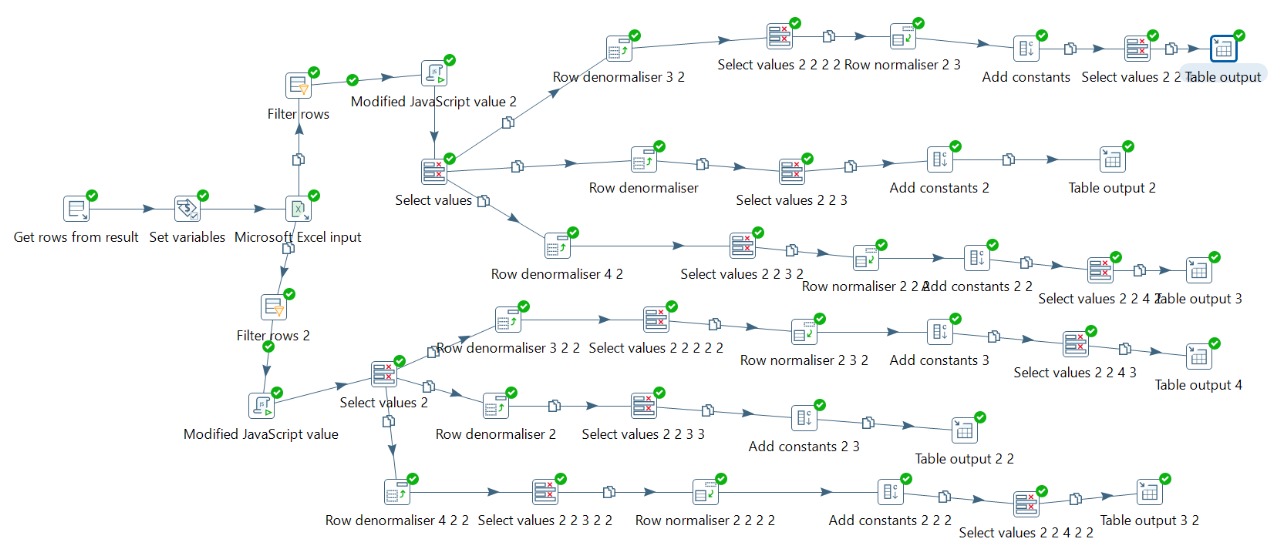


Figure 11: Transformation des données de Emploi

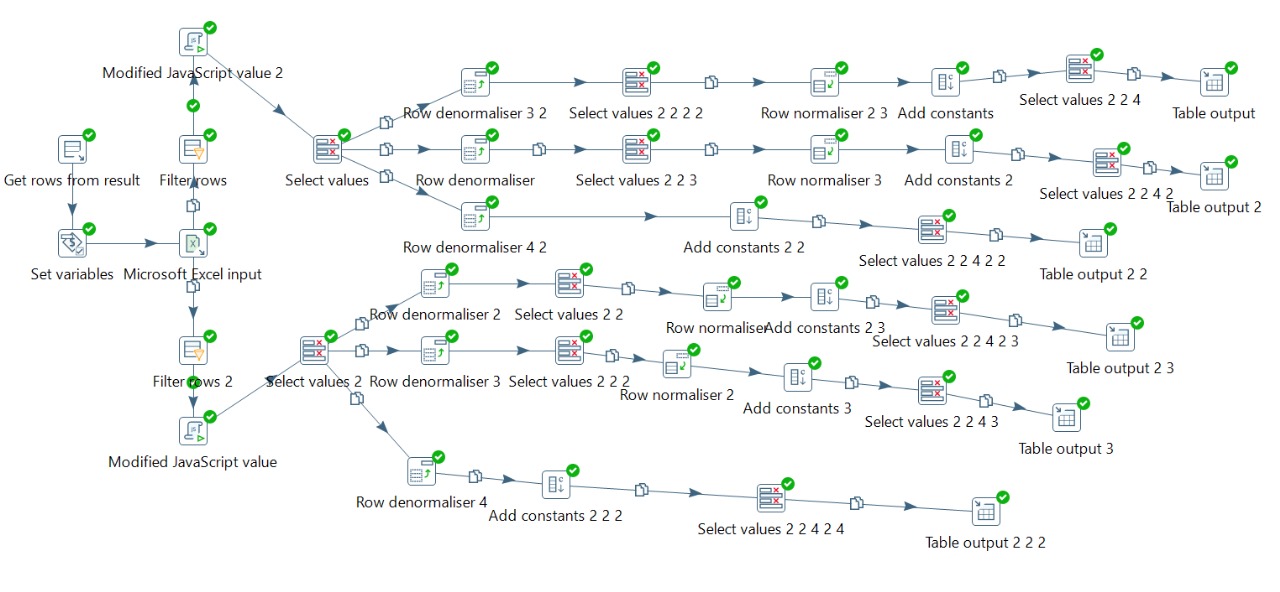


Figure 12: Transformation des données de Education

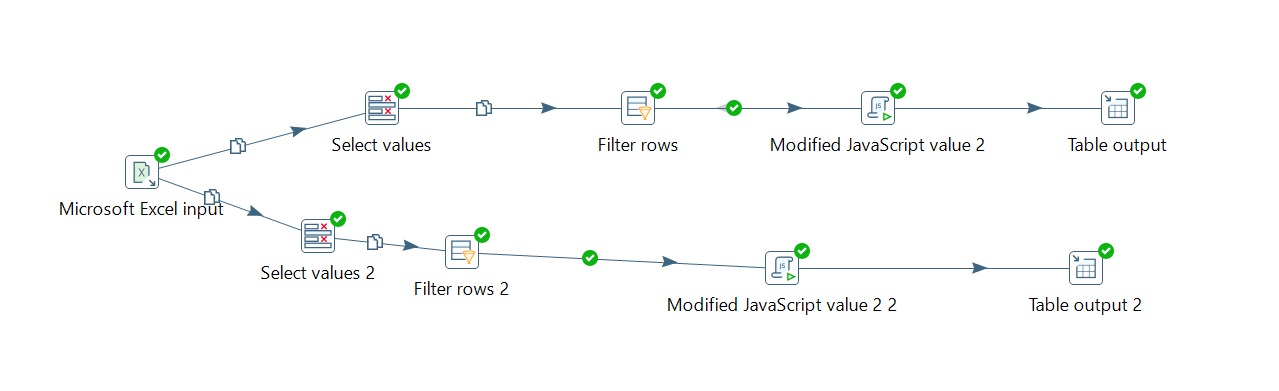


Figure 13: Transformation des données de santé

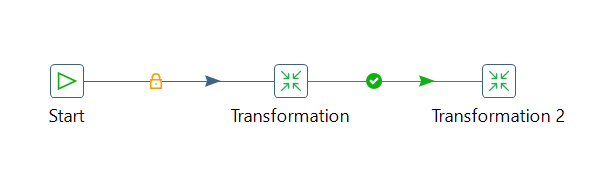


Figure 14: Job\_Automatisation\_transformation des fichier

Ce chapitre a présenté la modélisation multidimensionnelle en utilisant un schéma de constellation et a détaillé les étapes du processus ETL pour alimenter une base MySQL. L’approche a permis une organisation claire et efficace des données, prête pour des analyses décisionnelles.

# Chapitre3 : Création des visualisations interactives et automatisation

Une fois les données intégrées dans le data warehouse, des tableaux de bord interactifs ont été conçus pour visualiser les indicateurs clés et faciliter l’analyse des résultats. Ce chapitre regroupe la création des tableaux de bord, la stratégie de mise à jour des données, ainsi que les résultats obtenus et leur analyse. Ces éléments mettent en lumière l’impact des visualisations et leur utilité pour les décideurs et les chercheurs.

## Création des Tableaux de Bord

L’outil Power BI a été utilisé pour concevoir les tableaux de bord. Cet outil a permis de créer des visualisations interactives offrant une vue d’ensemble claire et intuitive des données.

Visualisations principales :

* Carte interactive :
  + Représente la répartition de la population par région et met en évidence les zones géographiques à forte ou faible densité.

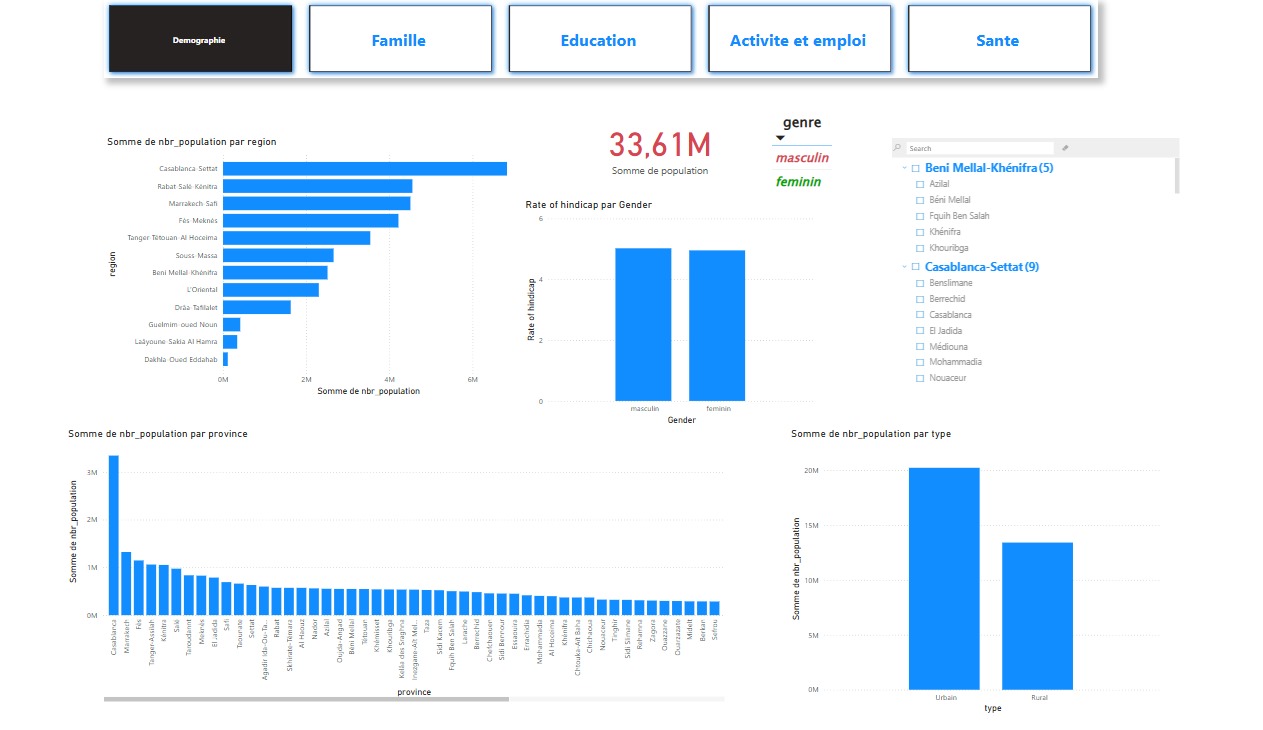


Figure 15:distrubution de population géographiquement

* Graphiques linéaires :
  + Illustrent l’évolution du nombre d’hôpitaux, de docteurs, de pharmacie, et de capacité litière au fil des années. Ces graphiques aident à identifier les tendances et les périodes critiques.

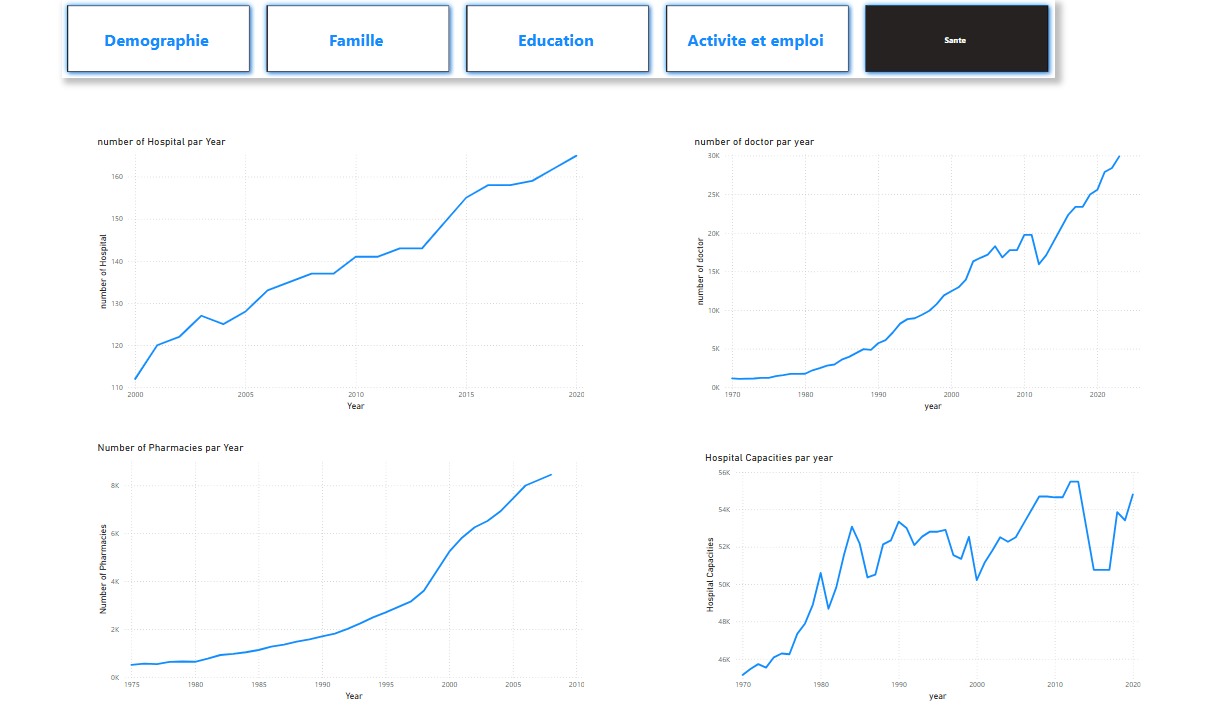


Figure 16:Evolution au niveau de sante

* Diagrammes circulaires :
  + Fournissent une analyse comparative des données par sexe et par milieu (urbain/rural), par exemple au niveau d’éducation dans le figure suivant :

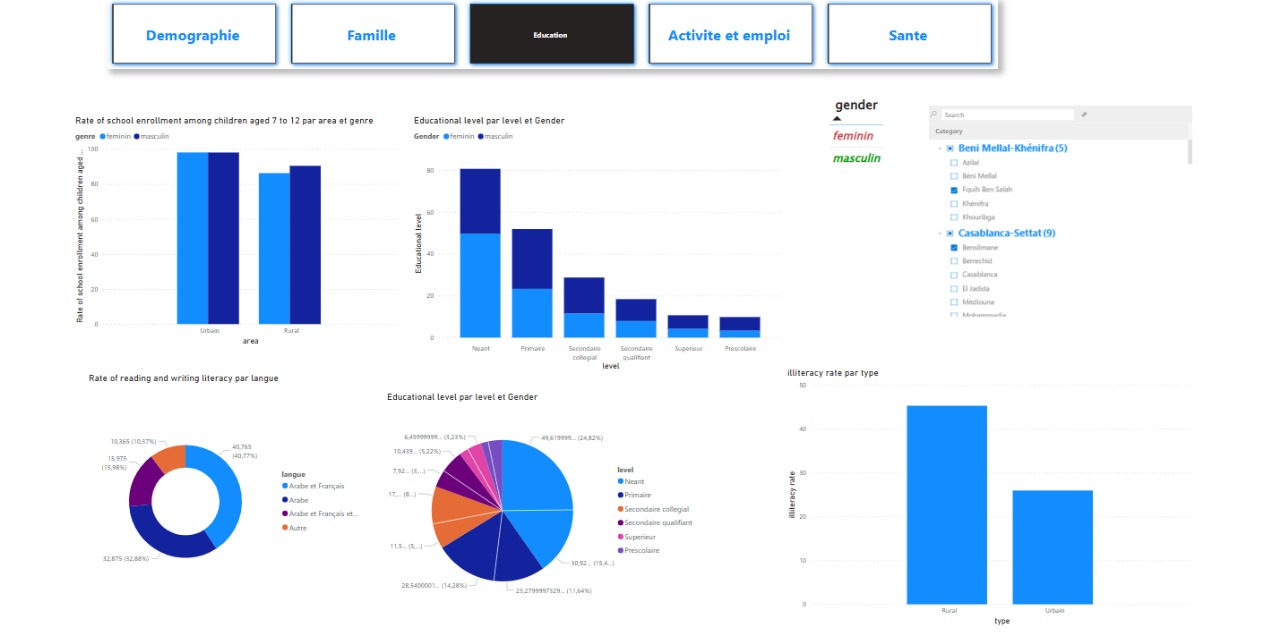


Figure 17:Visualisation, Analyse d'Education

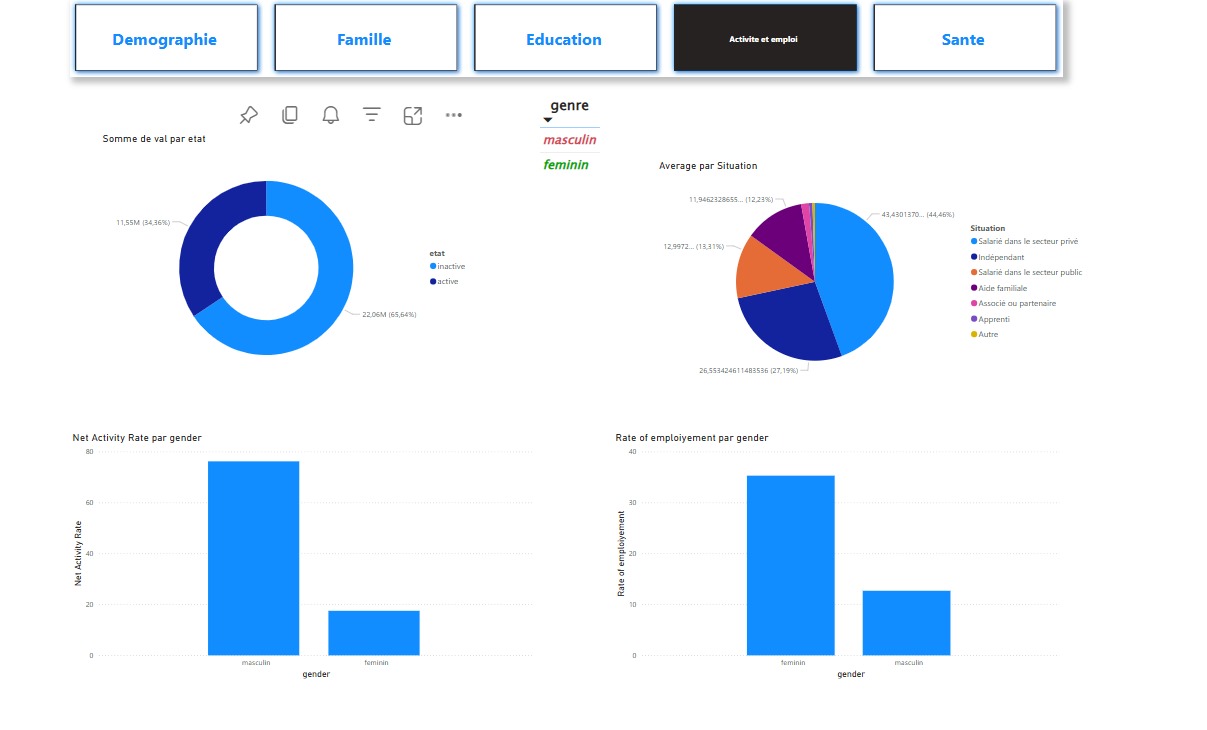
Des autre exemples :  


Figure 18:analyse au niveau d'Emploi

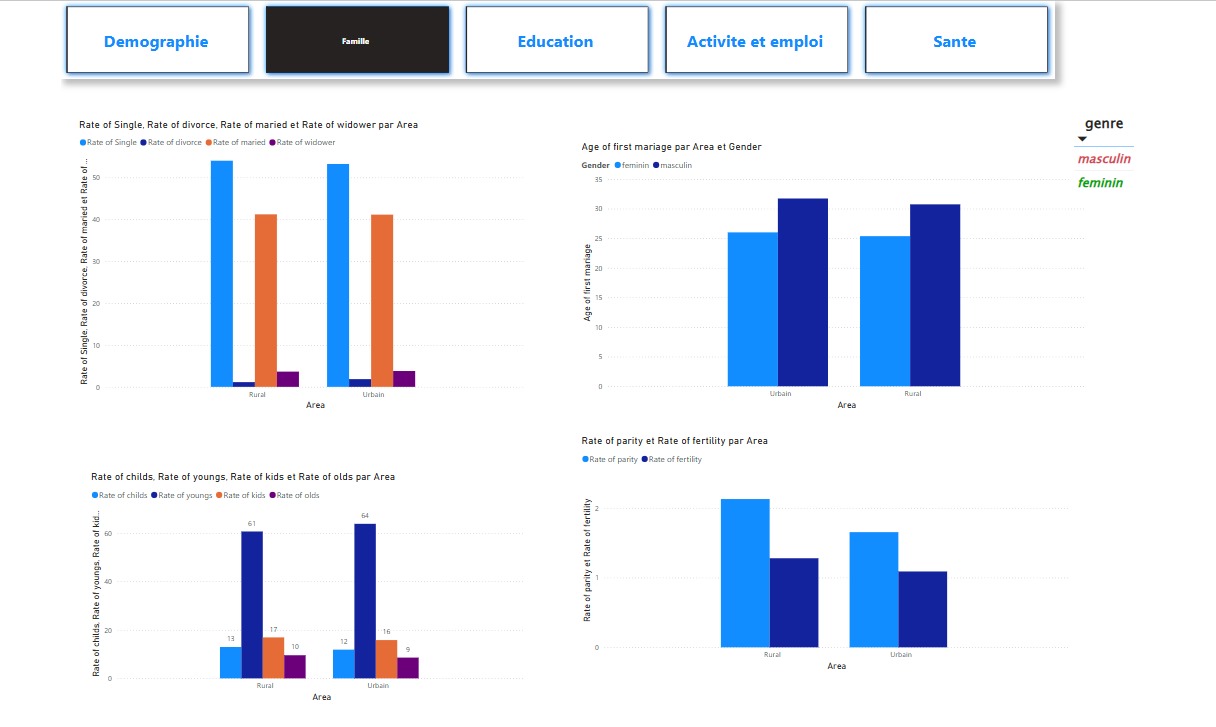


Figure 19: Analyse au niveau Famille

## Stratégie perspective de Mise à Jour Périodique

La mise à jour régulière des données est cruciale pour maintenir la fiabilité et l’actualité des tableaux de bord. Une stratégie d’automatisation et de planification sera mise en place pour gérer ce processus.

Automatisation :

* Des scripts Python pour extraire automatiquement les données des sources en utilisant le scraping des données et des jobs Pentaho (PDI) vont être développés pour transformer les données et les charger dans le data warehouse.

Planification :

* Des tâches planifiées, telles que les planificateurs de tâches, exécutent les scripts de scrapping et transformation à des intervalles définis.
* Des donnes du secteur de sante et emploi sont mis à jour chaque ans, alors les planificateurs de taches sera régler de s’exécuter chaque ans.

## Résultats Obtenus

Les tableaux de bord ont permis d’obtenir des insights précieux à partir des données disponibles.

Résultats principaux :

* Analyse régionale du chômage :
* Identification des régions avec les taux de chômage les plus élevés. Ces régions nécessitent des interventions politiques prioritaires.
* Disparités éducatives :
* Les visualisations mettent en évidence des écarts significatifs entre les milieux urbains et ruraux en matière d’accès à l’éducation et de taux de scolarisation.

## Analyse des Résultats

Les visualisations et les indicateurs clés offrent une base solide pour la prise de décision.

Utilité :

* Les décideurs politiques peuvent se baser sur les insights pour allouer des ressources et concevoir des stratégies ciblées.
* Les chercheurs peuvent utiliser les données pour approfondir l’étude des disparités régionales et sociales.

Limitations :

* Les résultats sont parfois impactés par des données manquantes ou obsolètes provenant des sources publiques par exemple province de Casablanca, Rabat ne contient pas des données rurales.
* Une amélioration continue des sources de données est nécessaire pour garantir une meilleure précision des analyses, aussi pour la tâche d’automatisation de mise a jours.

Les tableaux de bord créés à l’aide de Power BI offrent une vue d’ensemble interactive et claire des données, permettant une analyse approfondie et rapide. La stratégie de mise à jour garantit que les données restent à jour, rendant le système utile et fiable pour un large éventail d’utilisateurs.

# Conclusion

Ce projet a permis de concevoir un système de gestion et d’analyse de données, en intégrant des données provenant de diverses sources et en les organisant dans un schéma multidimensionnel. L'ETL a été utilisé pour transformer et charger ces données dans un système de manière automatisée, garantissant leur cohérence et fiabilité. Des visualisations interactives créées avec Power BI ont facilité l’analyse des indicateurs clés, tandis qu'une stratégie de mise à jour périodique a assuré l'actualité des données.

Les résultats ont montré l’efficacité du système pour analyser des données démographiques, éducatives et sociales, avec des insights utiles pour les décideurs. Cependant, certaines limitations liées à des données manquantes ont été identifiées.

Pour améliorer ce projet, il serait pertinent de travailler sur la qualité des données en réduisant les lacunes, et d'ajouter des dimensions économiques ou environnementales. Des modèles prédictifs pourraient également être intégrés pour anticiper les évolutions des indicateurs. Enfin, des visualisations plus interactives et une mise à jour en temps réel des données amélioreraient encore l'efficacité du système.