DÉCEMBRE 2021

DM

Management de projet : SCRUM

PROPOSÉ PAR

LA CLASSE 6 DE MASTER 2 DATA MANAGEMENT

ENSEIGNANT

OLIVIER MAMAVI

Paris School of Business

PARTIE 2: Construire une base de données

Chapitre 4: Base de données relationnelles

Projet

- le contexte

Une base de données relationnelle relie les informations entre elles au sein des bases de données. La base de données relationnelle stocke les données dans des tables, qui peuvent être accessibles et reconstruites de différentes manières, et qui sont elles-mêmes composées de lignes et de colonnes.

Dans le cadre de notre master nous avions pour objectif de construire une base de données. En effet nous devrions trouver des données , les nettoyer puis les relier afin de les rendre exploitables. Pour relier les données ou en d'autres termes les "jointer" il est nécessaire de passer par un système de gestion de base de données (SGBD). C'est le logiciel qui permet à un ordinateur de stocker, récupérer, ajouter, supprimer et modifier des données. Un SGBD gère tous les aspects primaires d'une base de données, y compris la gestion de la manipulation des données, comme l'authentification des utilisateurs, ainsi que l'insertion ou l'extraction des données. Un SGBD définit ce qu'on appelle le schéma de données ou la structure dans laquelle les données sont stockées.

- la mission

Comme abordé plus haut nous avions à notre disposition la base de données SIRENE qui se compose de plusieurs tables.

Deux tables élémentaires :

- TE_SIREN_ADMIN
- TE_SIREN_ADRESSE

Trois tables de référence :

- TR_CODE_EFFECTIF
- TR_NAF

- TR_NAP

Les tables élémentaires sont des tables volumineuses avec des informations pour chaque établissement.

Ces tables constituaient les premières informations que nous avions en notre possession. Par la suite nous avions recherché et trouvé deux autres tables:

-La table EGPRO:

Il s'agit de l'index de l'égalité professionnelle qui a été conçue pour faire progresser au sein des entreprises l'égalité salariale entre les femmes et les hommes. Il permet aux entreprises de mesurer, en toute transparence, les écarts de rémunération entre les sexes et de mettre en évidence leurs points de progression. Lorsque des disparités salariales sont constatées, des mesures de correction doivent être prises. L'index est calculé à partir de 5 indicateurs portant sur l'écart des rémunérations, promotions et augmentations, mais également sur le congé maternité et les plus hauts salaires de l'entreprise.

- La table Entreprise Ademe (REP-EMB) :

Le jeu de données présenté dans cette table comprend la liste des entreprises de la filière "Emballages ménagers" (EMB) inscrites dans SYDEREP. L'État a confié à l'ADEME la mise en place et la gestion des registres de déclarations annuelles obligatoires des metteurs sur le marché relevant des REP suivantes :

- Déchets des équipements électriques et électroniques (DEEE) ;
- Piles et accumulateurs (PA) ;
- Pneumatiques usagés (PU) ;
- Véhicules automobiles hors d'usage (VHU) ;
- Déchets d'éléments d'ameublement (DEA) ;
- Papiers graphiques (PAP) ;
- Emballages ménagers (EMB) ;
- Déchets issus de bateaux de plaisance ou de sport (DBPS).

Les entreprises se trouvant dans cette liste seront qualifiées d'entreprise "responsable" et celles qui n'y seront pas , seront invitées à la faire.

Notre mission ici sera d'identifier les entreprises à partir de leur numéro SIREN, d'affecter à chaque entreprise son indice d'égalité professionnelle. De même que si elle s'inscrit dans la bonne gestion de ses déchets. De façon à informer l'entreprise de sa situation et que des actions puissent en découler.

- le livrable

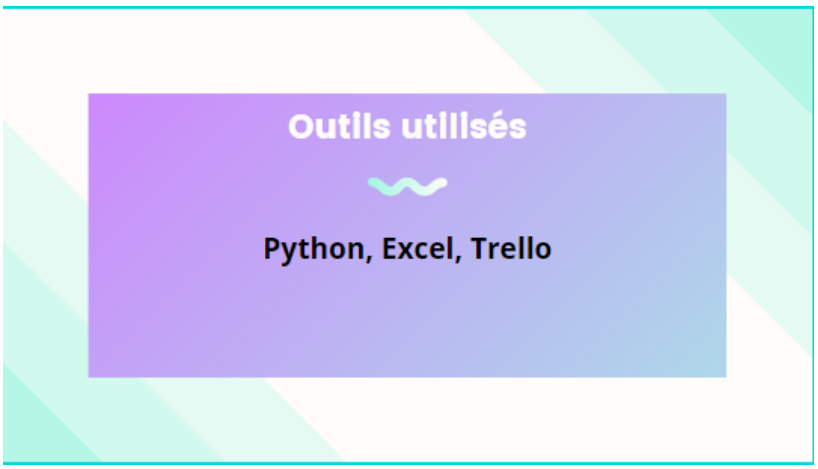
Le fichier final attendu sera sous un format csv regroupant toutes les informations en son sein nécessaire à l'identification et à la connaissance de la situation d'une entreprise.

Tâches réalisées

- l'organisation du projet :

Nous sommes un groupe de quatre filles qui constituons l'environnement. Les acteurs de ce sprint sont Sioutyne NGUY comme scrum master, Nadjelaa BENMESMOUDI comme project owner. Et tous les membres du groupe, Coumba DIALLO, Jocelyne TRAORE, Nadjelaa BENMESMOUDI et Sioutyne NGUY en qualité de développeurs.

Nous avions besoin à chaque sprint d'outils collaboratif et d'outils de gestion de base de données . En résumé les outils que nous avons le plus utilisées sont : Trello, Excel et Python



- la préparation du Sprint 1 :

Le premier sprint nous a permis de poser les bases de notre projet. Il avait pour objectif de nous aider à comprendre notre projet , de faire un inventaire de ce dont nous disposions. Ce que nous avions à faire, comment nous pourrions procéder et quel était le rendu attendu.

Chapitre 5: Collecter les données

Présentation du Sprint

- objectifs

Le premier sprint à eu lieu le 20 novembre 2021, pour le bon déroulement de notre projet plusieurs objectifs ont été déterminés, comme l'identification des sources de données, sélectionner des jeux de données, ainsi que la collecte des données qui nous semble en lien avec la responsabilité sociétale et environnementale des entreprises françaises.

- acteurs:

Les acteurs de ce sprint sont Sioutyne NGUY comme scrum master, Nadjelaa BENMESMOUDI comme project owner. Et tous les membres du groupe, Coumba DIALLO, Jocelyne TRAORE, Nadjelaa BENMESMOUDI et Sioutyne NGUY en qualité de développeurs.

- évènements

Comme le sujet de RSE est très large et vaste comme sujet, nous avons pris la décision de spécifier le type d'indicateurs que nous voudrions cibler dans notre projet, nous allons se focaliser sur les indicateurs RSE de type sociétale. De ce fait, nous avons construit la prochaine liste d'indicateurs :

- ➔ Nombre salarié Homme Femme
- ➔ Nombre de turnover
- ➔ Type de contrat
- ➔ Taux d'absentéisme
- ➔ Taux d'ancienneté
- ➔ Indice de satisfaction de qualité de vie et de santé du travail
- ➔ Taux d'activité selon le sexe
- ➔ Ecart salariale entre homme et femme
- ➔ Nombre de jours télétravail
- ➔ Mobilité des salariés

- support et outils utilisés

Outils de communication et gestion de projet: discord, messenger et trello

Outils de développement: python pour collecter les données, parsehub pour faire du scrapping

Tâches réalisées

- identifier des sources de données

Pour les sources de données, nous avons utilisé les deux sites open data, l'insee et data gouv

- sélectionner des jeux de données

Pour ce premier sprint, nous avons sélectionné trois jeux de données :

1- Egalité femmes-hommes :

Ce jeu de donnée traite le sujet d'inégalité entre femmes et hommes en terme d'espérance de vie, taux d'activité, taux de chômage, diplôme le plus élevé ainsi que l'écart de salaires entre les deux sexe.

2- Ecart Salariale entre les femmes et les hommes dans le secteur prive:

Ce deuxième jeu de données aborde le sujet d'écart de salaires entre les femmes et les hommes dans le secteur privé. Ce sont des données annuelles entre 2008 et 2018.

3- Salaire horaire moyen du mois de référence et rémunération brute totale dans les entreprises de 10 salariés et plus, y compris fonction publique d'État en 2014:

Ce dernier jeu de données de ce sprint présente le salaire horaire moyen du mois de référence et rémunération brute totale dans les entreprises de 10 salariés et plus, y compris fonction publique d'État en 2014.

- collecter les données

La collecte de données est un processus qui permet de recueillir de nombreux résultats dans des domaines divers et variés, les données que nous avons eu à partir des jeux de données seront ensuite préparées, organisées et analysées pour répondre à nos attentes en termes d'indicateurs sociétale RSE.

Résultats obtenus

Fichier sous format csv, issu des jointures entre les deux base de donnée 'Variables Siren' et 'Egalité hommes-femmes'

Difficultés rencontrées

Dans un premier temps, nous avons eu des difficultés sur où et comment trouver des jeux de données en lien avec notre sujet de projet. Puis, il n'était pas très évident d'obtenir, de collecter les données et de définir des indicateurs pertinents.

Un autre obstacle que nous avons rencontré est la visualisation des jeux de données, car ils ont une volumétrie importante qui ne pas être prise avec des outils simples comme excel.

Ce qu'il faut retenir

- Qu'est-ce que l'open data ?

Il s'agit de données auxquelles tout le monde peut accéder et que tout le monde peut utiliser et partager. On peut accéder aux données car elles sont disponibles en ligne d'une façon gratuite. On peut utiliser les données car elles sont disponibles sous une forme commune et lisible par des machines.L'open data doit être sous licence. Cette licence doit en permettre l'usage libre et autoriser les utilisateurs à transformer, combiner et partager ces données, même à des fins commerciales.

- Qu'est-ce que le webscraping ?

Le web scraping, appelé aussi harvesting, est une technique d'extraction du contenu de sites Web, via un script ou un programme, dans le but de le transformer pour permettre son utilisation dans un autre contexte, par exemple le référencement.

Chapitre 6: Préparer les données

Présentation du Sprint

- objectifs

L'objectif du sprint du 27 novembre 2021 était d'élaborer le modèle de la base de données, c'est-à-dire avoir un seul modèle de base de données qui regroupe toutes les informations pertinentes des entreprises.

- acteurs

Les acteurs de ce sprint sont Sioutyne NGUY comme scrum master, Nadjelaa BENMESMOUDI comme project owner.

Et tous les membres du groupe, Coumba DIALLO, Jocelyne TRAORE, Nadjelaa BENMESMOUDI et Sioutyne NGUY en qualité de développeurs.

- évènements

Nous avons sélectionnés plusieurs base de données comme :

1) Activité, emploi et chômage en 2020 et en séries longues

Cela nous permet de déterminer la population active et le taux d'activité par sexe en 2020.

La problème : difficulté à trouver des ressemblances entre les secteurs d'activité de cette table et celle de notre modèle de donnée

2) Masse salariale et assiette chômage partiel mensuelles du secteur privé, par région

Cela nous permet de comprendre ou se trouve la masse salariale (secteur public ou privé) ?

Le problème : difficulté à identifier les entreprises privée ou publique

3) Nombre d'établissements employeurs et effectifs salariés du secteur privé

Cela nous permet d'identifier les entreprises du secteur privé ce qui nous aide à la construction de la table

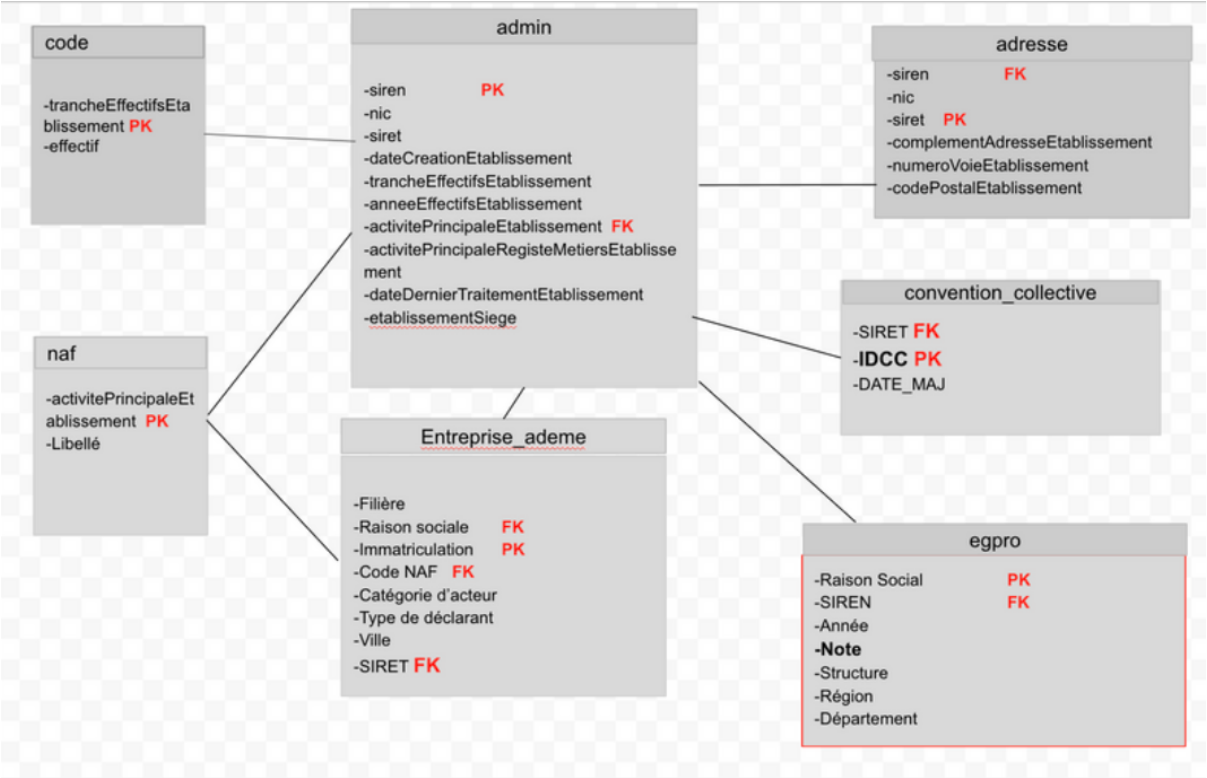
- support et outils utilisés

Outils de communication et gestion de projet: discord, messenger et trello

Autres outils : Python, Excel et Datastudio pour les problèmes de volumétrie

Tâches réalisées

- présenter les jeux de données



- décrire des variables

La table admin possède une clé primaire et étrangère

PK : Siren

FK : activitéPrincipaleEtablissement

La table code

PK : trancheEffectifsEtablissement

La table naf

PK : activitePrincipaleEtablissement

La table Entreprise_ademe

PK : Immatriculation

PK :

-Raison sociale

-Code NAF

SIRET

La table adresse

PK : Siret

FK : siren

La table convention_collective

PK : IDCC

FK : SIRET

La table egpro

PK : Raison social

FK : SIREN

Nous avons 7 tables.

- transformer des variables

Pour transformer les variables il faut nettoyer la base de donnée et enlever les doublons. Il faut également renommer les variables si nécessaire.

- indexer le jeu de données

Résultats obtenus

Les différentes tables nous permettent d'obtenir une table unique avec les différentes clefs primaires et secondaires nommée my_data.csv.

Difficultés rencontrées

Nous avons rencontré plusieurs difficultés :

1) Les données macro-économique : données par région, département plutôt que par entreprise
2) Les indicateurs pas assez précis
3) Séparation des libellés pour une meilleure compréhension
4) Difficulté à trouver des points en communs avec les entreprises pour créer les jointures

Ce qu'il faut retenir

- Qu'est-ce qu'une clé primaire ?

Dans une base de données, une clé primaire est un identifiant unique, en effet elle sert à identifier une ligne de manière unique.

Un champ est généralement affecté en tant que clé primaire, cela peut être un champ de texte ou un autre élément.

- Comment réaliser une jointure entre 2 tables de données ?

Pour réaliser une jointure en python, on utilise la librairie pandas (permet de manipuler et analyser les données en python).

La librairie pandas utilise des data frames qui correspondent aux tables en SQL.

On utilise donc la concaténation afin de faire les jointures entre les différentes tables. La fonction que nous avons utilisé pour concaténer nos dataframes est la fonction “merge”.

Chapitre 7: Organiser les données

Présentation du Sprint

- objectifs

  L'objectif de ce sprint était de de pouvoir nettoyer les tables, mettre en place toutes les jointures entre les différentes tables à notre disposition à l’aide des outils dédiés.

- acteurs

  Scrum master : Sioutyne Nguy

Product owner: Nadjelaa

Développeurs: Toute l'équipe

- évènements
    1. La réalisation d'un schéma relationnel pour nous permettre d'avoir l'architecture de notre base de données
    2. Procéder au nettoyage des différentes tables
    3. Jointures des différentes tables
    4. Obtention d'une base de données
- support et outils utilisés

  Pour les outils nous nous sommes servis de python pour nous permettre de concaténer nos tables

Tâches réalisées

- formater les variables (convertir) : dates, adresses, montant, ...

- nettoyer la base de données : doublons, valeurs extrêmes ou aberrantes, valeurs manquantes, ...

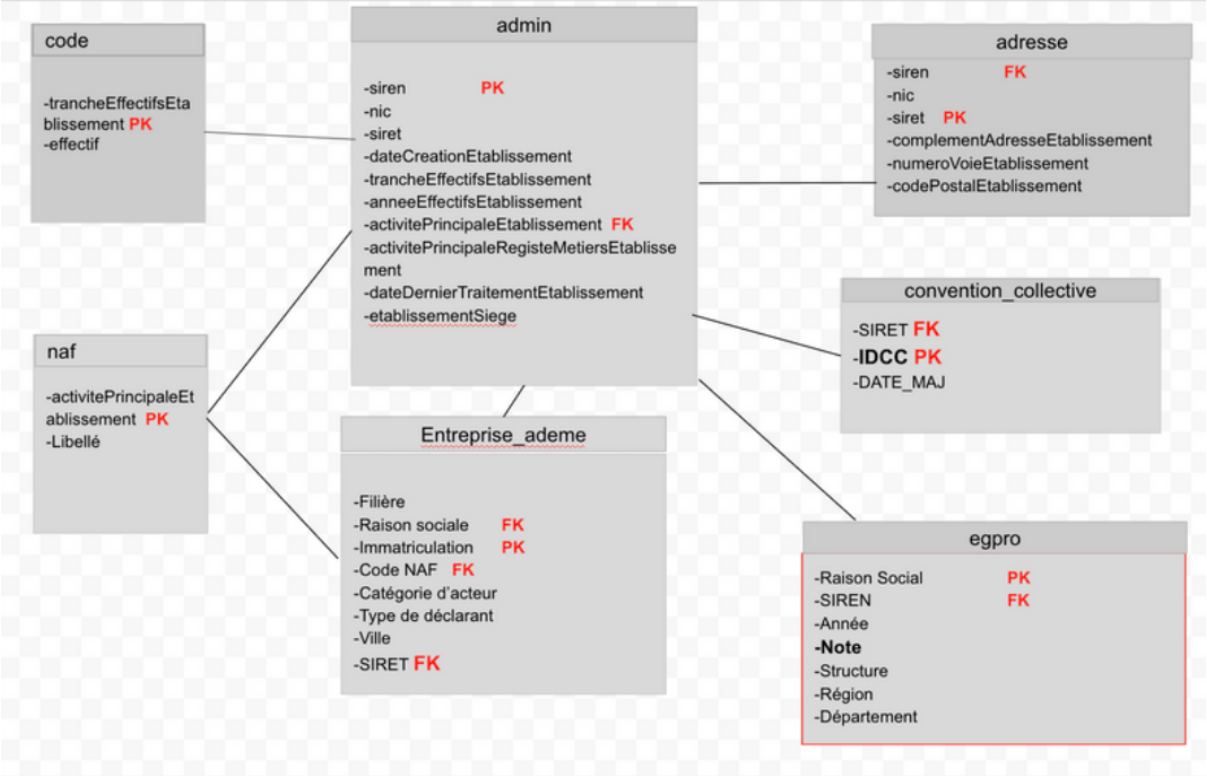
  Nettoyage des tables grâce à un **script python:**

```
df_paris = pd.read_csv('name.csv', sep=';')
```

- Pour penser l'organisation des variables dans des tables,Les tables dont les informations n'apportent une plus-value à notre base de données finale ont été supprimées toujours à l'aide de python

```
database = pd.merge(siret,adresse,
        how='left', left_on='parameter', right_on='id')
```

- Conception de la base de données



Résultats obtenus

- Ciblage des données
- Définition des jointures
- Réalisation du schéma relationnel

Difficultés rencontrées

- Impossible d'importer les tables : python avec pandas, phpmyadmin, postgres (point bloquant)
- Meilleure démarche technique à suivre (travailler par bd ou par table)
- Temps limité

Ce qu'il faut retenir

L'architecture de données est le processus qui permet de standardiser la façon dont les entreprises collectent, stockent, transforment, distribuent et utilisent les données. Le but est de fournir les données pertinentes aux personnes qui en ont besoin au moment opportun et de les aider à les interpréter.
L'architecte de données est chargé de définir la situation future, l'alignement au cours du développement et le suivi nécessaire pour s'assurer que la conception est faite selon les spécifications architecturales d'origine. Il considère ces systèmes de données comme une ressource stratégique de l'organisation, en les représentant indépendamment des processus des différentes unités qui les utilisent. Son principal but est de concevoir des structures de

données d'une manière organisée, fournissant ainsi une base pour la construction de systèmes d'information flexibles et intégrés.

Couche	Vision	Données	Intéressé
1	Portée / Contexte	Liste des choses importantes pour l'entreprise (domaines thématiques)	Responsable planification
2	Modèle Business / Conceptuel	Modèle sémantique ou Conceptuel / Enterprise Data Modeling (en)	Exploitant
3	Modèle Système / Logique	Entreprise / Modèle logique de données	Concepteur
4	Modèle Technologique / Physique	Modèle physique de données	Développeur
5	Représentations Détaillées	Base de données actuelle	Sous-traitant

Cadre Zachman

Chapitre 8: Présenter les données (data paper)

Présentation du Sprint

- objectifs

Réussir à avoir la base de données finale qui regroupe nos précédents jeux de données avec une architecture bien définie.

- acteurs

Les acteurs de ce sprint sont Sioutyne NGUY comme scrum master, Nadjelaa BENMESMOUDI comme project owner. Et tous les membres du groupe, Coumba DIALLO, Jocelyne TRAORE, Nadjelaa BENMESMOUDI et Sioutyne NGUY en qualité de développeurs.

- évènements

Finaliser la base de donnée finale et pouvoir la stocker sur un espace de stockage

- support et outils utilisés

Python, excel et trello

Tâches réalisées

- Résumé
- Contexte et objectifs
- démarche et organisation de la base de données
- Description des variables
- Exploitation et usages

Difficultés rencontrées

Ce qu'il faut retenir

- Qu'est-ce qu'un data paper ?

Le data paper est une publication qui décrit un jeu de données scientifiques brutes (data, dataset), notamment à l'aide d'informations précises, appelées métadonnées (metadata).

Il est publié sous forme d'articles examinés par les pairs dans une revue publiant généralement différentes formes d'articles dont des data papers ou dans un data journal, c'est-à-dire une revue contenant exclusivement des data papers.

Le data paper comprend deux parties :

- l'ensemble des fichiers des données (data files) accessibles directement ou via un entrepôt de données. Ces données sont en libre accès ou en restriction d'accès temporaire ;
- la partie descriptive, c'est le data paper proprement dit. Cette partie explique le contexte d'obtention des données, les présente et en démontre la fiabilité.

Le data paper a pour objectif d'informer la communauté scientifique de l'existence et de la disponibilité d'un jeu de données qui est déposé dans un entrepôt de données. Dans notre cas nous utiliserons comme canal Gitea :

Il valorise les données en exposant leur potentiel pour des utilisations et projets futurs. Il facilite leur réutilisation en mettant en évidence la qualité des données et des procédures, ainsi que la rigueur scientifique de l'étude. Il apporte de la visibilité aux données, les rend plus facilement repérables et citables par d'autres études. Ainsi notre travail pourra être utile à d'autres personnes en recherche d'information tel que l'index d'égalité professionnelle par exemple sur une entreprise ou un établissement français à partir du numéro SIREN. Le data paper met aussi en valeur ses auteurs en tant que créateurs de données.

Le contenu du data paper est le suivant :

- Titre et auteurs:

Datasets de données RSE sociétaux

par Diallo Coumba, Pamidjo Jocelyne, Benmesmoudi Nadjelaa, Nguy Sioutyne

- Résumé

La responsabilité sociétale et environnementale des entreprises comporte de nombreux enjeux pour l'entreprise qui souhaite intégrer sa stratégie globale, liés aux objectifs de développement durable.
La RSE comporte un volet social important qui se traduit par des objectifs en matière de conditions de travail, de bien-être et de motivations des collaborateurs. Les actions mises en place visent à améliorer l'environnement de travail et augmenter la satisfaction des collaborateurs. Dans notre étude, nous nous sommes intéressés plus précisément aux inégalités hommes-femmes dans le milieu professionnel en termes d'effectif, rémunération et plein d'autres enjeux.

Contexte et objectifs

Durant ce cours, nous avons dû créer des équipes et définir le rôle de chacun afin de gérer un projet avec la méthode agile. Ainsi, nous voilà une équipe de 5 personnes entièrement féminine et dont le nom est Totally Sprint.
Ce nom fait référence aux Totally Spies, un dessin animé de notre enfance ou des filles sauvent le monde grâce à des missions ! Fantastique n'est-ce pas ?

A partir de métadonnées fournies par notre professeur, données portant sur la RSE des entreprises. Nous avions comme objectif de chercher et sélectionner des indicateurs pertinents afin de mettre en avant la responsabilité sociétale des entreprises. De plus, grâce à notre étude, nous avons la possibilité de mettre en avant les bons comme les mauvais comportements des entreprises. Notre étude pourra aider les entreprises à s'améliorer et à évoluer pour le bien de la société.

Etant une équipe de filles, le sujet de l'égalité salariale hommes femmes au travail nous paraissait le plus important comme nous y sommes actuellement et dans le futur impactés. De plus, nous sommes tous concernés par l'environnement et par conséquent l'empreinte carbone des entreprises.

- Démarche et organisation de la base de données

La récolte de données RSE sociétaux est souvent liée à la recherche d'indicateurs pariétaux ou encore à ceux affiliés à des indicateurs comme l'égalité, à une géolocalisation des structures ou encore à des données liées à des critères beaucoup plus spécifiques.Ainsi dans le cadre de la mise en place de nos jeux de données,

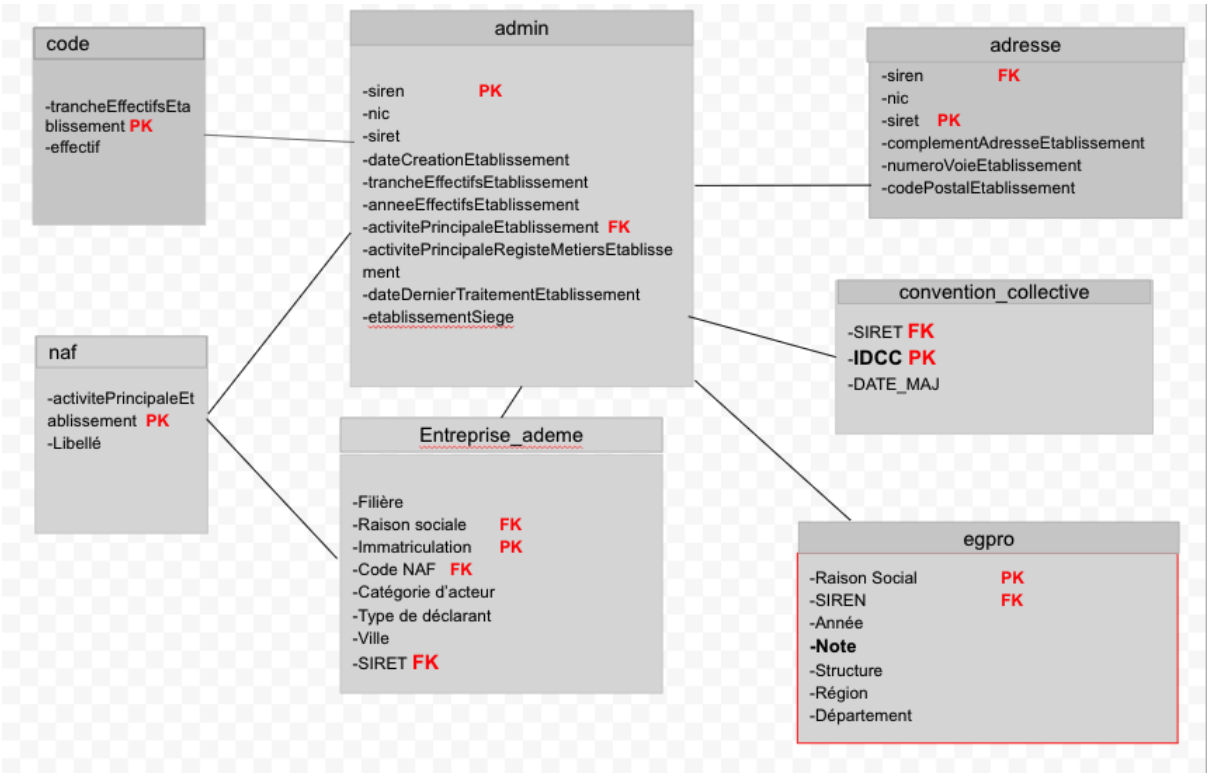
certains jeux de données ont été jugés pertinents pour nos recherches pour aboutir à l'établissement d'un jeu de données final.Ainsi plusieurs étapes successives sont notées dans ce processus:

- Collecte des jeux de données sur différentes sources notamment Datagouv et INSEE
- L'identification des tables et le choix des différents attributs des tables
- Le nettoyage des datasets
- La mise en place d'un schéma relationnel pour avoir une structuration du dataset final
- Procéder à une jointure de toutes les différentes tables
- Obtention de la table finale
- Mise à disposition des annexes sur Gitea

Organisation du dataset

Le dataset final est une jointure des différents datasets que nous avons eu à collecter:
Adresse

- Description des variables et des tables



Variable	Type	Source/Ingénierie
Siren	numeric	admin
nic	numeric	admin
siret	numeric	admin
dateCreationEtablissement	date	admin
trancheEffectifsEtablissement	numeric	admin
anneeEffectifsEtablissement	integer	admin
activitePrincipaleEtablissement	character	naf
Libellé	character	admin
activitePrincipaleREgistreMetiersEtablissement	character	admin
dateDernierTraitementEtablissement	date	admin
etablissementSiege	character	admin
complementAdresseEtablissemen	character	adresse

t		
numeroVoieEtablissement	numeric	adresse
CodePostaleEtablissement	numeric	adresse
trancheEffectifsEtablissement	numeric	code
effectif	numeric	code
Raison Social	character	egpro
Année	numeric	egpro
Note	numeric	egpro
Structure	character	egpro
Région	character	egpro
Département	character	egpro
IDCC	numeric	convention_collective
DATE_MAJ	date	convention_collective
Filière	character	Entreprise_ademe
Immatriculation	numeric	Entreprise_ademe
Catégorie d'acteur	character	Entreprise_ademe
Type de déclarant	character	Entreprise_ademe
Ville	character	Entreprise_ademe

- Description de chaque table source:

1.egpro :

Description

L'Index de l'égalité professionnelle a été conçue pour faire progresser au sein des entreprises l'égalité salariale entre les femmes et les hommes.
Il permet aux entreprises de mesurer, en toute transparence, les écarts de rémunération entre les sexes et de mettre en évidence leurs points de
progression. Lorsque des disparités salariales sont constatées, des mesures de correction doivent être prises.

L'index est calculé à partir de 5 indicateurs portant sur l'écart des rémunérations, promotions et augmentations, mais également sur le congé maternité et les plus hauts salaires de l'entreprise.

2.convention_collective :

Description

Ce jeu de données présente les conventions collectives déclarées par entreprise (SIRET)
Ce jeu de données est exploité dans le projet Code du travail numérique
Il est issu de la DSN et mis à jour ponctuellement avec plusieurs mois de retard.
Le colonne IDCC représente le numéro de convention collective, laquelle peut être consultée via https://beta.legifrance.gouv.fr/recherche

3.Entreprise_ademe (REP-EMB):

Description

Le jeu de données
Le jeu de données présenté ici comprend la liste des entreprises de la filière "Emballages ménagers" (EMB) inscrites dans SYDEREP.
Date des données : juin 2020 (en cours d'actualisation)
Règles d'actualisation des données
Les données concernent la liste des producteurs enregistrés à un instant T sur SYDEREP, soit par l'intermédiaire d'un éco-organisme, soit en direct dans le cas d'un système individuel. Ces listes ne reflètent pas les producteurs qui ont effectué une déclaration sur SYDEREP.
L'actualisation de cette liste est mise à disposition une fois par an sur https://data.ademe.fr.
Généralités sur les filières REP
L'État a confié à l'ADEME la mise en place et la gestion des registres de déclarations annuelles obligatoires des metteurs sur le marché relevant des REP suivantes :
Déchets des équipements électriques et électroniques (DEEE) ;
Piles et accumulateurs (PA) ;
Pneumatiques usagés (PU) ;
Véhicules automobiles hors d'usage (VHU) ;
Déchets d'éléments d'ameublement (DEA) ;
Papiers graphiques (PAP) ;
Emballages ménagers (EMB) ;
Déchets issus de bateaux de plaisance ou de sport (DBPS).
Cette liste va s'enrichir au fur et à mesure de l'entrée en vigueur de la loi AGEC avec :
Au 01/01/2022, les filières : Produits du tabac, Lubrifiants (HU), Articles de bricolage et de jardin (ABJ), Articles de sport et loisirs (ASL), Jouets, Produits et matériaux du secteur de la construction du bâtiment (PMCB),
Au 01/01/2023, la filière Emballages issus de la restauration,
Au 01/01/2024, les filières Gommes à mâcher et Textiles sanitaires à usage unique,

Au 01/01/2025, la filière Engins de pêche
Ces registres sont gérés dans l'application SYDEREP, permettant aux éco-organismes ou aux producteurs en système individuel, le renseignement direct de leurs données. Ces registres ayant pour objet la mesure de la performance globale au niveau national de chacune des filières REP, l'ADEME établit et publie annuellement, via sa Librairie un état des lieux des filières avec un certain nombre d'indicateurs agglomérés, complétés d'éléments d'analyses qualitatives provenant des acteurs de chaque filière.

- Exploitation et usages

Nous disposions dès le départ d'informations sur les entreprises à partir d'une base de données comportant le numéro SIREN / SIRET, le code NAF , les tranches d'effectif et plusieurs éléments liées à l'identification des entreprises et des établissements.
L'enjeu pour nous était donc de pouvoir exploiter ses informations en trouvant d'autres données qu'on pourrait croiser avec celles que nous avons déjà afin de déduire des informations pertinentes et utiles . La mise en commun des données est essentielle pour permettre de développer une stratégie data à haute valeur ajoutée pour l'entreprise.
En croisant ces différentes données, il est possible d'accéder à des informations plus pointues et surtout plus riches sur leur marché, leurs métiers et surtout leur situation face aux problématiques actuelles.

Dans notre cas nous nous sommes beaucoup intéressés aux sujets liées à l'égalité professionnelle , l'écart des rémunérations, promotions et augmentations entre homme et femme. Et surtout la gestion des déchets générés par les produits fabriqués ou mis sur le marché (responsabilité élargie du producteur REP).
Notre base de données permettra donc à connaître pour chaque entreprise sa situation face au traitement entre les femmes et les hommes en termes d'accès à l'emploi, à la formation, à la mobilité et à la promotion ou en termes d'égalité salariale. Si elle arrive à gérer correctement ses résidus de fabrication de produits, si elle participe à réduire les déchets qui polluent la planète et si elle est exempte de toute réprimande.
Le plus important est en fait de mettre en évidence les points de progression sur lesquels agir. La finalité est qu'à partir de nos extractions qu'on puisse avertir une entreprise sur sa façon de fonctionner et qu'elle puisse prendre des actions afin d'éviter de mauvaises notations

Notre dépôt sur Gitea ICI