| **Type de données** | **Fournisseur direct** | **Source réelle** | **Ownership** | **Remarques** |
| --- | --- | --- | --- | --- |
| **Prépaiements** | DFIN | SGRF DAT / ALM | SGRF/ALM | Fournies pour RACER |
| **Origination volume** | DFIN | SGRF DAT / ALM | SGRF/ALM | Fournies pour RACER |
| **Pertes (PES, PED)** | RISQ CRE / ERM | RISQ CRE / ERM | RISQ | Données certifiées par RISQ |
| **Encours sains/défaut** | DFIN (extraction SAFIR) | SAFIR (RISQ) | RISQ | DFIN n’est **pas propriétaire** de ces données |
| **Provisions** | DFIN (extraction SAFIR) | SAFIR (RISQ) | RISQ | À valider avec RISQ CRE/ERM |

**Classe DataQuality**

La classe DataQuality est conçue pour regrouper l’ensemble des contrôles qualité automatisés sur les données collectées et nettoyées dans les classes précédentes. Ces contrôles sont essentiels pour garantir la robustesse et la fiabilité des indicateurs produits.

**Méthodes principales**

* **check\_missing\_values(threshold\_dict)**  
  Calcule le taux de valeurs manquantes par colonne et le compare à un seuil critique défini. Les colonnes dépassant ce seuil sont signalées dans le rapport.
* **check\_duplicates()**  
  Identifie les doublons dans le jeu de données. Les lignes dupliquées sont extraites et stockées pour analyse.
* **check\_outliers(threshold\_dict)**  
  Détecte les valeurs extrêmes ou incohérentes via des règles statistiques ou métiers (ex. : seuils, percentiles). Le résultat est conservé dans un rapport spécifique.
* **sanity\_check(columns\_to\_sanity\_check)**  
  Vérifie des règles métier sur certaines colonnes critiques (par exemple : une date de défaut ne peut pas être postérieure à la date d’analyse). En cas d’incohérence majeure, une exception DataQualityError est levée.
* **Génération des rapports : generate\_dq\_report(output\_folder)**

Convertit les résultats des contrôles en fichiers .csv.

Les fichiers sont écrits dans un sous-dossier nommé selon le type de données (ex. sanity\_check\_pd, sanity\_check\_lgd), dans un dossier géré Dataiku.

Utilise les bibliothèques pandas ou Spark selon la source des données.