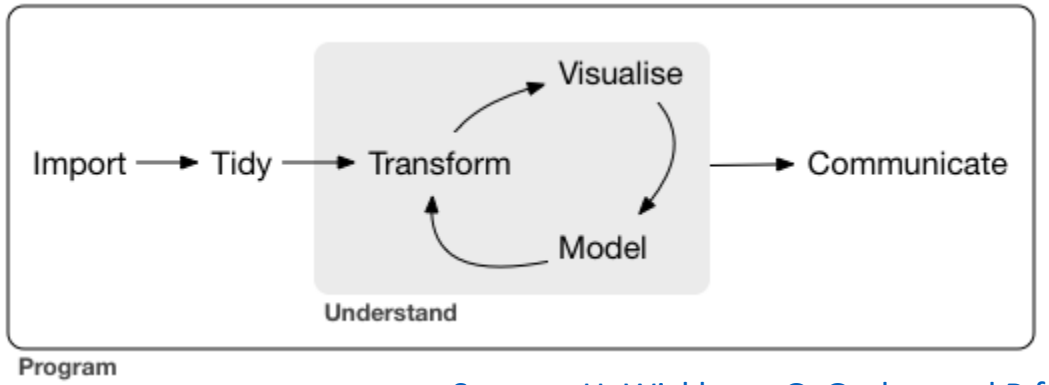
SAE 2.04 Exploitation d'une base de données

Un exemple de mini-projet relevant des « Data sciences »

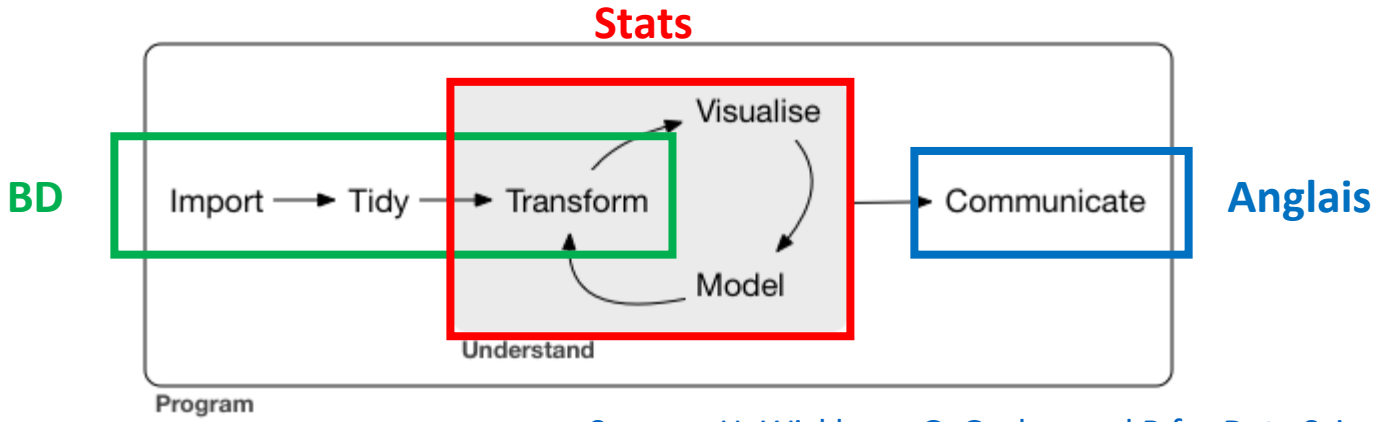


Source : H. Wickham, G. Grolemund R for Data Science

Mise en situation :
Vous êtes deux membres d'une petite association sensible à l'accès libre aux données et souhaitant informer le consommateur sur le nutriscore [VIDEO].

Objectif du projet :
A partir d'une base de données réelle disponible en *opendata* et renseignée par des contributions libres (donc mal remplie, et contenant plus d'informations que nécessaire etc.), réaliser une étude simple contenant des graphiques appropriés pour répondre à une question précise, et rédigée en anglais.

SAE 2.04 Exploitation d'une base de données



Source : H. Wickham, G. Grolemund R for Data Science

Calendrier du projet :

- Semaine 1 [Ma 02/04 - Je 04/04] : Exploration et nettoyage des données sous postgreSQL (3x2h)
- Semaine 1 [Ve 05/04] : Production des graphiques (3h)
- Semaine 2 [Lu 08/04] : Rédaction en anglais (2h)

Organisation du travail :

- Projet réalisé en binômes (intra-groupe A,B,C,D,E)
- Des évaluations complémentaires individuelles.

Gestion de projet : cadrage du projet

Nature du travail : Cadrage du projet

- Cadrage du projet
- Analyse des contraintes et identification des risques
- Identification du processus de travail

Données entrées : transparents de présentation et sujet
Données sorties : Questionnaires envoyé par binôme

Organisation du temps :

- En autonomie
- Dates de rendu : Contraintes et risques (**Lundi 08 Avril**) en.pdf sur Chamilo

Evaluation : sur la base d'un questionnaire

- Cadrage du projet : Identification des contraintes et des risques / Gestion des risques

SQL : exploration et nettoyage des données

Données en entrée : une vue personnelle en lecture seule vers une table commune où ont été importées les données originelles (plus de 3 millions de lignes, env. 7 Gio).

Données en sortie : un script psql contenant :

- Les requêtes SQL et commandes psql permettant, à partir de cette vue, de générer un fichier CSV ne contenant que les données utiles, nettoyées.

Organisation du temps :

- Un enseignant circule entre les salles pour répondre aux questions
- Script psql à rendre **Jeudi 04 Avril 12h** (midi).
- Dès réception de tous les fichiers : partage d'un fichier CSV pour chaque sujet.

Evaluation :

- Script psql : qualité, originalité.
- Exercice lors du contrôle sur table de R2.06 : explication de la solution que vous avez mise en œuvre.

Statistiques : analyse

Données en entrée : un fichier CSV ne contenant que les données utiles.

Données en sortie : au minimum 5 tableaux et/ou graphiques produits via R avec les commentaires, utiliser Knit pour produire un fichier PDF

Organisation du temps :

- un enseignant circule dans les salles pour répondre aux questions : lire le tableau sur le calcul du Nutriscore, faire quelques statistiques simples sur vos données pour présenter les variables retenues et éventuellement repérer des valeurs aberrantes, chercher les couleurs du Nutriscore et enfin, produire des graphiques et/ou des tableaux ainsi que les commentaires .
- **Date de rendu :** au plus tard le **Mardi 09/04/2024 à 23h59.**

Evaluation : pertinence des tableaux, des graphiques et des commentaires.
A rendre : le fichier PDF

Anglais : exposition

Données en entrée : Les tableaux et/ou graphiques que vous aurez produits via R.

Données en sortie : Le fichier PDF rédigé en anglais analysant les graphiques et faisant ressortir les informations pertinentes.

Organisation du temps :

- Un enseignant circule dans les salles pour répondre aux questions.
- **Date de rendu : Mardi 09/04 au soir (23h59 dernier délai)**

Evaluation : pertinence de l'analyse, qualité et pertinence de la langue utilisée, structuration claire du dossier