Introduction sur les données manquantes

Marie Vaugoyeau

20 February 2024

Table of contents

1 Définitions **1**

2 Type de données manquantes **2**

3 Conséquences des valeurs manquantes **2**

4 Identifier les valeurs manquantes **2**

5 L'analyse descriptive **3**

6 Traitement des valeurs manquantes **3**

7 En savoir un peu plus sur moi **4**

1 Définitions

Les données manquantes sont les données qui ne sont pas présentes.
La donnée peut-être remplacée dans le tableau par :

- `NA`
- Une autre valeur dépendante des données ou de la personne qui s'en ai occupée : `0`, `NO`, `999`...

> **Note**
>
> Quelques soit le cas, il existent plusieurs origines aux données manquantes.

2 Type de données manquantes

Les données manquantes, représentées par `NA` ou autre peuvent avoir plusieurs origine :

- La donnée **n'est pas compatible**. *Par exemple, une personne rentre du texte au lieu d'un numéro de téléphone.* Dans ce cas le **système ne prends pas en charge** la réponse et la qualifie en `NA` pour `Not Applicable`

- La donnée **n'existe pas**. *Par exemple la personne n'a pas de numéro de téléphone*, dans ce cas, le système la qualifie de `NA` pour `Not Available`

- La donnée **existe mais n'a pas été communiquées**. *Par exemple la personne a refusé de donner son numéro*, dans ce cas, le système la qualifie de `NA` pour `Not Answer`

Dans tous les cas, la **seule information transmise** est que la **données n'est pas disponible**.
Il n'est pas toujours possible de cerner l'origine du problème mais cela n'empêche pas d'agir.
Il faut commencer par se demander ce que **signifie cette absence** et **comment elle va impacter** notre système.

3 Conséquences des valeurs manquantes

- Perte d'information : Si la donnée peut-être retrouvée ou remplacée, pourquoi s'en empecher ?

- Erreur dans la généralisation : Si beaucoup de donnnées sont manquantes et que les conclusions se basent uniquement sur celles présentent, **est-ce que cela représente vraiment la réalité ?**

- Comportement de certains modèles stats

4 Identifier les valeurs manquantes

Pour savoir comment agir, il faut commencer par quantifier et localiser les valeurs manquantes.

Une réalisation simple est l'utilisation de la fonciton `summary()` du package `{base}`.

Le package `{naniar}` est spécialement adapté à la visualisation des données manquantes.

Visualisation du nombre ou de la proportion de données manquantes grâce aux fonctions `gg_miss_var()`.

La fonction `gg_miss_upset()` permet de représenter sur un graphique les variables qui ont des données manquantes et le lien entre les colonnes.

La fonction `geom_miss_point()` permet de visualiser les valeurs manquantes sur les nuage de poin.

Et il existe d'autres fonctions :

- `vis_miss()` du package `{visdata}`
- `md.pattern()` du package `{mice}`

> ⚠ Détection des valeurs manquantes
>
> Les données manquantes peuvent avoir été remplacées par d'autres. Il est possible de les détecter grâce à l'**analyse descriptive**.

5 L'analyse descriptive

L'analyse descriptive a pour but d'analyser les variables pour connaître la nature des données mais aussi identifier les valeurs extrêmes (à ne pas confondre avec aberrentes).

Utilisation de fonctions rapide comme :

- `skim()` du package `{skimr}`
- `dfSummary()` du package `{summarytools}`
- `create_report()` du package `{DataExplorer}`

6 Traitement des valeurs manquantes

- Remplacer la donnée manquante par :
  - La vraie valeur s'il est possible de la retrouver.
  - Une valeur de remplacement :
    - Déterminée à partir des autres données de la variables : *moyenne*, *médiane*, *minimum*, *maximum*...

    * Modélisée à partir des autres variables grâce à *une régression linéaire*, *une ACP*...

- Ne pas remplacer la données mais garder le `NA`
- Supprimer la ligne ou la colonne concernée. **Cette solution est la moins envisageable et ne doit être mise en place que si les deux autres ne sont pas possibles.**

7 En savoir un peu plus sur moi

Bonjour,

Je suis Marie Vaugoyeau et je suis disponible pour des **missions en freelance** d'**accompagnement à la formation** à R et à l'analyse de données et/ou en **programmation** (reprise de scripts, bonnes pratiques de codage, développement de package). Ayant un **bagage recherche en écologie**, j'ai accompagné plusieurs chercheuses en biologie dans leurs analyses de données mais je suis ouverte à d'autres domaines.

Vous pouvez retrouver mes offres ici.

En plus de mes missions de consulting je diffuse mes savoirs en R et analyse de données sur plusieurs plateformes :

- J'ai écrit un **livre** aux éditions ENI
- Tous les mois je fais un **live sur Twitch** pour parler d'un package de R, d'une analyse
- Je rédige une **newsletter** de manière irrégulière pour parler de mes **inspirations** et transmettre **des trucs et astuces sur R**. Pour s'y inscrire, c'est par là. J'ai aussi un **blog** sur lequel vous pourrez retrouver une version de cet article.

Pour en savoir encore un peu plus sur moi, il y a LinkedIn et pour retrouver tous ces liens et plus encore, c'est ici

N'hésitez pas à me contacter sur marie.vaugoyeau@gmail.com !

Bonne journée

Marie


Chercheuse-re-s, entreprises,
je vous accompagne dans
la valorisation de VOS données