Rendu

November 14, 2024

ROCHETTE LEVEQUE

1 Chargement des Données

Nous importons le dataset pour obtenir une vue d'ensemble de ses caractéristiques et de ses variables.

[1]:
```
data <- read.csv("../data/support2.csv")
head(data)
```

A data.frame: 6 × 47

	age <dbl>	death <int>	sex <chr>	hospdead <int>	slos <int>	d.time <int>	dzgroup <chr>	dzcl <ch
1	62.84998	0	male	0	5	2029	Lung Cancer	Can
2	60.33899	1	female	1	4	4	Cirrhosis	CO
3	52.74698	1	female	0	17	47	Cirrhosis	CO
4	42.38498	1	female	0	3	133	Lung Cancer	Can
5	79.88495	0	female	0	16	2029	ARF/MOSF w/Sepsis	ARI
6	93.01599	1	male	1	4	4	Coma	Cor

2 Description des Variables

Le dataset contient `nrow(data)` observations et `ncol(data)` variables. Voici une description des principales variables :

1. **Variables Démographiques**
   - `age` : Âge du patient (numérique).
   - `sex` : Genre du patient (catégorique - "male" ou "female").
   - `edu` : Niveau d'éducation, avec des valeurs manquantes.
   - `income` : Niveau de revenu, également avec un nombre important de valeurs manquantes.
2. **État de santé et résultats**
   - `death` et `hospdead` : Indicateurs de mortalité (binaire).
   - `slos` : Durée de séjour à l'hôpital.
   - `d.time` : Durée jusqu'au décès ou au dernier suivi.
   - `dzgroup` et `dzclass` : Variables catégorielles spécifiant le groupe de maladie et la classe.

3. **Mesures cliniques**
    - `sps`, `aps`, `scoma` : Scores représentant l'état de santé et le niveau de coma.
    - Signes vitaux tels que `meanbp` (pression artérielle moyenne), `hrt` (fréquence cardiaque), `resp` (taux de respiration) et `temp` (température).
    - Résultats de tests de laboratoire : `pafi`, `alb`, `bili`, `crea`, `sod`, `ph`, `glucose`, `bun` et `urine`.
4. **Fonctionnalité et suivi**
    - `adlp` et `adls` : Scores des activités de la vie quotidienne, avec de nombreuses valeurs manquantes.
    - `sfdm2` : Statut de suivi, également avec quelques valeurs manquantes.

3 Analyse des valeurs manquantes

Nous examinons le nombre de valeurs manquantes pour chaque variable.

[2]:
```
missing_data_summary <- sapply(data, function(x) sum(is.na(x)))
missing_data_summary <- sort(missing_data_summary[missing_data_summary > 0],␣
↪decreasing = TRUE)
missing_data_summary
```

adlp 5641 **urine** 4862 **glucose** 4500 **bun** 4352 **totmcst** 3475 **alb** 3372 **adls** 2867 **bili** 2601 **pafi** 2325 **ph** 2284 **prg2m** 1649 **edu** 1634 **prg6m** 1633 **totcst** 888 **wblc** 212 **charges** 172 **avtisst** 82 **crea** 67 **dnrday** 30 **scoma** 1 **sps** 1 **aps** 1 **surv2m** 1 **surv6m** 1 **meanbp** 1 **hrt** 1 **resp** 1 **temp** 1 **sod** 1

Les variables avec des valeurs manquantes significatives sont :

- `edu` (niveau d'éducation),
- `income`,
- `pafi` (pression partielle en oxygène artériel),
- `alb` (albumine),
- `glucose`,
- `bun` (azote uréique sanguin),
- `urine`,
- `adlp` et `adls` (scores d'activités de la vie quotidienne).

4 Analyse des valeurs aberrantes

Pour détecter les valeurs aberrantes, nous analysons les valeurs situées en dehors de 1,5 fois l'écart interquartile (IQR).

[3]:
```r
find_outliers <- function(series) {
  q1 <- quantile(series, 0.25, na.rm = TRUE)
  q3 <- quantile(series, 0.75, na.rm = TRUE)
  iqr <- q3 - q1
  lower_bound <- q1 - 1.5 * iqr
  upper_bound <- q3 + 1.5 * iqr
  return(sum(series < lower_bound | series > upper_bound, na.rm = TRUE))
}

outliers_summary <- sapply(data[, sapply(data, is.numeric)], find_outliers)
outliers_summary <- sort(outliers_summary[outliers_summary > 0], decreasing =␣
 ↪TRUE)
outliers_summary
```

scoma 1955 **diabetes** 1778 **hday** 1543 **crea** 987 **bili** 926 **charges** 912 **dnrday** 799 **slos** 768 **totcst** 749 **totmcst** 495 **wblc** 399 **resp** 313 **surv2m** 307 **dementia** 296 **sps** 283 **glucose** 272 **d.time** 267 **bun** 267 **ph** 260 **sod** 256 **edu** 199 **aps** 178 **adlp** 149 **urine** 92 **pafi** 90 **age** 56 **avtisst** 43 **hrt** 40 **num.co** 25 **alb** 15 **temp** 14 **meanbp** 6

5 Interprétation des valeurs aberrantes

Les variables présentant le plus grand nombre de valeurs aberrantes sont :

- `slos` : Durée de séjour, avec des valeurs extrêmes suggérant que certains patients ont eu de très longues hospitalisations.
- `charges` et `totcst` : Frais hospitaliers avec des valeurs particulièrement élevées pour certains patients.
- Les indicateurs de santé, tels que `bili`, `crea`, `glucose` et `bun`, présentent également des valeurs extrêmes, ce qui peut refléter des anomalies dans les conditions de santé de certains patients.

6 Conclusion

Cette analyse descriptive permet d'identifier les variables avec des valeurs manquantes et des valeurs aberrantes, fournissant une base pour des traitements ultérieurs des données et des analyses plus approfondies.