missing-value-imputation

Max Lang

10/27/2022

handle_missing

This function gives several options to imputing or remove missing values in a data.frame.

Inputs

- df: A dataframe
- choice: Either "remove" or "impute", to define the missing value handling
- missing: The encoding of missing values

```
# Helper Function
getmode <- function(v, na.rm = TRUE) {
    if (na.rm) v <- na.omit(v)

    uniqv <- unique(v)
    uniqv[which.max(tabulate(match(v, uniqv)))]
}

handle_missing <- function(df, choice, missing = NA) {
    assertDataFrame(df, types = c("character", "numeric"))
    assert(testCharacter(missing, any.missing = TRUE, len = 1),
        testInt(missing),
        combine = "or"
    )
    assertChoice(choice, choices = c("remove", "impute"))

    df[df == missing] <- NA  # recode missing values as NA's

    if (choice == "remove") {
        df <- na.omit(df)
    } else {
        for (i in colnames(df)) {
            col <- df[[i]]
            if (is.numeric(col)) {
                df[[i]][is.na(col)] <- median(col, na.rm = TRUE)
            } else {
                df[[i]][is.na(col)] <- getmode(col)
            }
        }
    }

    return(df)
}
```

Worked example

```
df <- data.frame(v1= rnorm(10), v2= rnorm(10), v3= rnorm(10))
# Sample random observations that miss
df[sample(10,5),1] <- NA
df[sample(10,3),2] <- NA
df[sample(10,2),3] <- NA

handle_missing(df, choice = "remove", missing = NA)
```

```
##           v1        v2          v3
## 4  0.2066644 0.2257745 -1.54154186
## 5 -0.1668887 1.7783667 -0.08895065
## 6  1.4285754 0.3439763  0.05438805
```

```
handle_missing(df, choice = "impute", missing = NA)
```

```
##            v1         v2          v3
## 1  -0.1668887  0.3439763  0.69724915
## 2  -0.3142404  0.3439763 -0.88568325
## 3  -0.1668887  1.2034607 -0.01728130
## 4   0.2066644  0.2257745 -1.54154186
## 5  -0.1668887  1.7783667 -0.08895065
## 6   1.4285754  0.3439763  0.05438805
## 7  -1.1201029  0.5244085 -0.01728130
## 8  -0.1668887 -0.5713019  0.12950413
## 9  -0.1668887  0.3439763 -0.37018278
## 10 -0.1668887 -0.2627566  1.87730451
```