Pre_Process Dataset Information

Diane Hoang

2024-05-13

```
## TEAM HIRE ME NOW ##

## Step 1 Import the Dataset

# Load the readr package
library(readr)

# Use read_csv to import the CSV file
df <- read_csv("/Users/dianehoang/Documents/Drug_overdose_death_rates__by_drug_type__sex__age__race__and
```

```
## Rows: 6228 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): INDICATOR, PANEL, UNIT, STUB_NAME, STUB_LABEL, AGE, FLAG
## dbl (8): PANEL_NUM, UNIT_NUM, STUB_NAME_NUM, STUB_LABEL_NUM, YEAR, YEAR_NUM,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## Step 2 Make a format that is easy to review all the unique variables and then decide how we want to

# Function to store column names, unique values, data type, and number of null values in a data frame
get_column_info_df <- function(df) {
  column_info <- data.frame(Column_Name = character(),
                            Unique_Values = character(),
                            Data_Type = character(),
                            Num_Null = integer(),
                            stringsAsFactors = FALSE)
  for (col in names(df)) {
    if (is.numeric(df[[col]])) {
      unique_values <- paste(unique(df[[col]]), collapse = ", ")
      data_type <- "Numeric"
    } else {
      unique_values <- paste(unique(as.character(df[[col]])), collapse = ", ")
      data_type <- "Character"
    }
    num_null <- sum(is.na(df[[col]]))
    column_info[nrow(column_info) + 1, ] <- list(col, unique_values, data_type, num_null)
  }
  return(column_info)
}
```

```r
# Call the function to get column info as a data frame
column_info_df <- get_column_info_df(df)

# Print the data frame
print(column_info_df)
```

```
##       Column_Name
## 1       INDICATOR
## 2           PANEL
## 3       PANEL_NUM
## 4            UNIT
## 5        UNIT_NUM
## 6       STUB_NAME
## 7   STUB_NAME_NUM
## 8      STUB_LABEL
## 9  STUB_LABEL_NUM
## 10           YEAR
## 11       YEAR_NUM
## 12            AGE
## 13        AGE_NUM
## 14       ESTIMATE
## 15           FLAG
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14 6.1, 6.2, 6.8, 8.2, 8.9, 9.4, 10.1, 11.5, 11.9, 12.3, 13.2, 13.1, 13.8, 14.7, 16.3, 19.8, 21.7, 8
## 15
##    Data_Type Num_Null
## 1  Character        0
## 2  Character        0
## 3    Numeric        0
## 4  Character        0
## 5    Numeric        0
## 6  Character        0
## 7    Numeric        0
## 8  Character        0
## 9    Numeric        0
## 10   Numeric        0
## 11   Numeric        0
## 12 Character        0
## 13   Numeric        0
## 14   Numeric     1111
```

```
## 15 Character     5117
```

```
## Saved the results in Excel and sent to team to show in the Data understanding section for review and


library(openxlsx)

# Define the file path for the Excel file
file_path <- "/Users/dianehoang/Documents/Team 7/column_info_Drug Overdose.xlsx"

# Write the data frame to an Excel file
write.xlsx(column_info_df, file_path, row.names = FALSE)
```

```
## Warning: Please use 'rowNames' instead of 'row.names'
```

```
# Print a message indicating the file has been saved
cat("Excel file saved successfully to:", file_path, "\n")
```

```
## Excel file saved successfully to: /Users/dianehoang/Documents/Team 7/column_info_Drug Overdose.xlsx
```