data preparation 6 - 10 / 3 / 2023

Objective

The purpose of this notebook is data preparation, namely: rename variables in English, remove special characters or whitespaces, create simple categorical variables to better represent the information, build a single `.csv` file.

The code is written as a `function()` and applied to all source files `*.xlsx`.

```
prepare_for_analysis <- function(input_file,
                                 selected_cols,
                                 new_column_names){

  df <- read_excel(input_file)  %>%

    # remove unnecessary columns
    select(all_of(selected_cols)) %>%

    # rename columns
    setNames(new_column_names) %>%

    #replace spaces with _ in column names
    rename_with(~ gsub(" ", "_", .), everything()) %>%

    #remove special characters such as ô, â ...
    mutate(across(where(is.character), ~ gsub("[^a-zA-Z0-9\\s]", " ", .))) %>%

    #set appropriate format for Day and get weekday
    mutate(day = as.Date(day, format = "%d-%m-%Y")) %>%
    mutate(weekday = as.numeric(format(day, "%u"))) %>%

    # encode direction as inbound or outbound
    mutate(direction = case_when(
      grepl("di ritorno da altro Comune", type) ~ "inbound",
      grepl("diretto in altro Comune", type) ~ "outbound",
      TRUE ~ "--")) %>%

    # encode type of traveller
    mutate(res_trav = case_when(
      grepl("Residente", type) ~ "resident",
      grepl("Viaggiatore", type) ~ "traveller",
      TRUE ~ "--"))

    return(df)
}
```

select and rename columns:

```r
selected_cols <-  c(
  "Giorno", "Comune di partenza", "Comune di arrivo", "Tipologia viaggiatore", "Viaggi",    "00-03", "03

new_column_names <- c(
  "day", "origin", "destination", "type", "n",
  "00_03", "03_06", "06_09", "09_12", "12_15", "15_18", "18_21", "21_24")

input_file <- './data/EXPORT FVG _ 6-10 MARZO 2023.xlsx'

df <- input_file %>%
  prepare_for_analysis(selected_cols, new_column_names)

print(df)
```

```
## # A tibble: 257,985 x 16
##    day        origin     destination type      n `00_03` `03_06` `06_09` `09_12`
##    <date>     <chr>      <chr>       <chr> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 2023-03-06 Porcia     Pordenone   Resi~  1876   0.005   0.006   0.493   0.148
##  2 2023-03-06 Pordenone  Porcia      Resi~  1794   0.013   0       0.017   0.096
##  3 2023-03-06 Pordenone  Cordenons   Resi~  1695   0.004   0       0.025   0.132
##  4 2023-03-06 Cordenons  Pordenone   Resi~  1667   0.009   0       0.451   0.162
##  5 2023-03-06 Tavagnacco Udine       Resi~  1425   0.003   0.02    0.51    0.155
##  6 2023-03-06 Udine      Tavagnacco  Resi~  1366   0.016   0.003   0.021   0.093
##  7 2023-03-06 Trieste    Muggia      Resi~  1325   0.011   0.027   0.235   0.222
##  8 2023-03-06 Muggia     Trieste     Resi~  1302   0.017   0.017   0.056   0.134
##  9 2023-03-06 Sacile     Fontanafre~ Resi~  1267   0.024   0.003   0.078   0.143
## 10 2023-03-06 Fontanafr~ Sacile      Resi~  1248   0.012   0.018   0.352   0.171
## # i 257,975 more rows
## # i 7 more variables: `12_15` <dbl>, `15_18` <dbl>, `18_21` <dbl>,
## #   `21_24` <dbl>, weekday <dbl>, direction <chr>, res_trav <chr>
```

```r
# filter relevant rows
print("BEROFRE filtering:")
```

```
## [1] "BEROFRE filtering:"
```

```r
table(df$type)
```

```
##
##             Residente Comunale di ritorno da altro Comune di Provincia di residenza
##                                                                                34628
##   Residente Comunale di ritorno da altro Comune di Regione  esterno alla Provincia
##                                                                                23002
##                Residente Comunale diretto in altro Comune di Provincia di residenza
##                                                                                35418
##      Residente Comunale diretto in altro Comune di Regione  esterno alla Provincia
##                                                                                24462
##         Viaggiatore Estero diretto in qualsiasi altro Comune  interno alla Regione
##                                                                                18993
##      Viaggiatore Nazionale diretto in qualsiasi altro Comune  interno alla Regione
##                                                                                39014
##           Viaggiatore Provinciale diretto in altro Comune di Provincia di residenza
##                                                                                37063
## Viaggiatore Provinciale diretto in altro Comune di Regione  esterno alla Provincia
##                                                                                13853
```

```
##             Viaggiatore Regionale diretto in altro Comune di Provincia di residenza
##                                                                                14733
##   Viaggiatore Regionale diretto in altro Comune di Regione  esterno alla Provincia
##                                                                                16819
```

```
#df <- df %>% filter(str_detect(type, "^Residente Comunale"))
df <- df %>% filter(res_trav == "resident")

print("AFTER filtering:")
```

```
## [1] "AFTER filtering:"
```

```
table(df$type)
```

```
##
##           Residente Comunale di ritorno da altro Comune di Provincia di residenza
##                                                                                34628
## Residente Comunale di ritorno da altro Comune di Regione  esterno alla Provincia
##                                                                                23002
##              Residente Comunale diretto in altro Comune di Provincia di residenza
##                                                                                35418
##    Residente Comunale diretto in altro Comune di Regione  esterno alla Provincia
##                                                                                24462
```

```
df %>% write.csv("dati_6_10_marzo_2023.csv")
```