Milestone # 2

Moliehi Mokete and Bongekile Nkosi

2022-10-01

This is a team assignment; each team should complete and turn in a PDF created from an Rmd via Github. Please include code and output for the following components:

Description of data set 1. What is the data source? (1-2 sentences on where the data is coming from, dates included, etc.) **Data sources**

monkey_pox data source is from European Center for Disease Prevention and Control(ECDC)

pop_denominator data source is from European commission

census_stat data source is from European census statistics 2011

world_country_region data source is from European census

2. How does the data set relate to the group problem statement and question?

The data set is going to help us understand how monkey pox case rates may differ by region and various demographic factors,additionally, the data set will allow us to determine if there is a relationship between certain demographics and monkey pox case rates.

Import statement

NOTE: Please use data sets available in the PHW251 Project Data github repo Links to an external site. (this is important to make sure everyone is using the same data sets)

Use appropriate import function and package based on the type of file(done)

Loading the library to be used in this assignment

```
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command ’timedatectl’
## had status 1

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(janitor)
```

```
##
## Attaching package: ’janitor’

## The following objects are masked from ’package:stats’:
##
##     chisq.test, fisher.test
```

```
library(purrr)
library(stringr)
```

Utilize function arguments to control relevant components (i.e. change column types, column names, missing values, etc.)

Importing data from git repositories

```
file_path1<-"https://raw.githubusercontent.com/PHW290/phw251_projectdata/main/euro_mpx_cases.csv"
monkey_pox <-read_csv(file_path1,na = c("", "NA", "*", "n/a"))%>% clean_names()
```

```
## Rows: 2987 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): CountryExp, CountryCode, Source
## dbl  (1): ConfCases
## date (1): DateRep
##
## 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.
```

```
file_path2<-"https://raw.githubusercontent.com/PHW290/phw251_projectdata/main/euro_pop_denominators.csv
pop_denominator<- read.csv(file_path2,na = c("", "NA", "*", "n/a"))%>% clean_names()%>% rename(country_


file_path3<-"https://raw.githubusercontent.com/PHW290/phw251_projectdata/main/euro_census_stats.csv"
census_stats <- read.csv(file_path3,na = c("", "NA", "*", "n/a"))%>% clean_names()


file_path4<-"https://raw.githubusercontent.com/PHW290/phw251_projectdata/main/world_country_regions.csv
world_country_region <- read.csv(file_path4,na = c("", "NA", "*", "n/a"))%>% clean_names()
```

Document the import process(working on it)

Identify data types for 5+ data elements/columns/variables Utilize functions or resources in RStudio to determine the types of each data element (i.e. character, numeric, factor)

```
map(monkey_pox, class)
```

```
## $date_rep
## [1] "Date"
##
## $country_exp
## [1] "character"
##
## $country_code
## [1] "character"
##
## $source
## [1] "character"
##
## $conf_cases
## [1] "numeric"
```

```
map(pop_denominator,class)
```

```
## $dataflow
## [1] "character"
##
## $last_update
## [1] "character"
##
## $freq
## [1] "character"
##
## $indic_de
## [1] "character"
##
## $country_code
## [1] "character"
##
## $time_period
## [1] "integer"
##
## $obs_value
## [1] "integer"
##
## $obs_flag
## [1] "character"
```

```
map(world_country_region,class)
```

```
## $name
## [1] "character"
##
## $alpha_2
## [1] "character"
##
## $alpha_3
## [1] "character"
##
## $country_code
## [1] "integer"
##
## $iso_3166_2
## [1] "character"
##
## $region
## [1] "character"
##
## $sub_region
## [1] "character"
##
## $intermediate_region
## [1] "character"
##
## $region_code
## [1] "integer"
##
## $sub_region_code
## [1] "integer"
```

```
map(census_stats,class)
```

```
## $country_code
## [1] "character"
##
## $sex
## [1] "character"
##
## $age
## [1] "character"
##
## $cas
## [1] "character"
##
## $edu
## [1] "character"
##
## $time
## [1] "integer"
##
## $flags
## [1] "character"
##
## $footnotes
## [1] "character"
##
## $res_pop
## [1] "character"
```

```
##
## $pop
## [1] "integer"
```

Identify 5+ data elements required for your specified scenario. If <5 elements are required to complete the analysis, please choose additional variables of interest in the data set to explore in this milestone.

confirmed cases,date reported, sub_region, time period, age, education, sex, employment status

Identify the desired type/format for each variable—will you need to convert any columns to numeric or another type?

education from character to string, date reported from date to integer (month), Provide a basic description of the 5+ data elements Numeric: mean, median, range

```
summary(monkey_pox$conf_cases)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   0.000   5.715   1.000 655.000
```

```
summary(pop_denominator$time_period)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    2011    2013    2016    2016    2019    2022
```

Character: unique values/categories

```
x<-unique(world_country_region$sub_region)
x
```

```
##  [1] "Southern Asia"                   "Northern Europe"                
##  [3] "Southern Europe"                 "Northern Africa"                
##  [5] "Polynesia"                       "Sub-Saharan Africa"             
##  [7] "Latin America and the Caribbean" "Western Asia"                   
##  [9] "Australia and New Zealand"       "Western Europe"                 
## [11] "Eastern Europe"                  "Northern America"               
## [13] "South-eastern Asia"              "Eastern Asia"                   
## [15] "Melanesia"                       "Micronesia"                     
## [17] "Central Asia"
```

```
z<-unique(census_stats$age)
z
```

```
## [1] "Y_GE85" "Y_LT15" "Y15-29" "Y30-49" "Y50-64" "Y65-84"
```

```
w<- unique(census_stats$edu)
w
```

```
## [1] "ED1"  "ED2"  "ED3"  "ED4"  "ED5"  "ED6"  "NAP"  "NONE" "UNK"
```

```
y<- unique(census_stats$sex)
y
```

```
## [1] "F" "M"
```

Or any other descriptives that will be useful to the analysis