

WikiData Processing: European Biographical Data

Julia Ulziisaikhan

6/15/2022

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.5      v purrr   0.3.4
## v tibble  3.1.5      v dplyr   1.0.7
## v tidyr   1.1.4      v stringr 1.4.0
## v readr   2.0.2      v forcats 0.5.1
```

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

```
library(ggplot2)
library(readxl)
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

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

```
path <- "C:/Users/ulzii/OneDrive/Documents/spring22/tuttle/wikidata_analysis"
setwd(path)
```

Data Processing

Our lifespan column is calculated from subtracting the person's birth year from the death year. We exclude persons with lifespans less than 0.

```
#functions
blank2na <- function(field){
  output <- ifelse(field=="", NA, field)
  return(output)
}

raw2df <- function(dat){
  df <-
    dat %>%
    clean_names() %>%
    mutate(age = yod - yob,
           occupation_label=blank2na(occupation_label),
           citizen_label=blank2na(citizen_label),
           birthcountry_label = blank2na(birthcountry_label)
    ) %>%
    filter(!is.na(age)) %>%
    filter(age>=0) %>%
    group_by(item) %>%
    slice(1) %>%
    ungroup() %>%
    relocate(age, .after=yod) %>%
    relocate(item, .after=birthcountry_label) %>%
    arrange(yob)
  return(df)
}
```

```
#Load in data
france <- read.csv("data/france_query.csv")
germany <- read.csv("data/germany_query.csv")
russia <- read.csv("data/russia_query.csv")
italy <- read.csv("data/italy_query.csv")
uk <- read.csv("data/uk_query.csv")
spain <- read.csv("data/spain_query.csv")
greece <- read.csv("data/greece_query.csv")
portugal <- read.csv("data/portugal_query.csv")
```

```
df <- rbind(raw2df(france),
            raw2df(germany),
            raw2df(russia),
            raw2df(italy),
            raw2df(uk),
            raw2df(spain),
            raw2df(greece),
            raw2df(portugal))
```

```
colnames(raw2df(france))
```

```
## [1] "item_label"      "yob"      "yod"
## [4] "age"             "occupation_label" "citizen_label"
## [7] "birthcountry_label" "item"
```

We remove the country of citizenship column and instead assign country of birth as the person's associated country. `id` refers to the WikiData link for the given person's entry.

```
df$citizen_label <- NULL
colnames(df) <- c("name", "birth", "death", "lifespan", "occupation", "country", "id")
df$dataset <- "wikidata"
head(df)
```

```
## # A tibble: 6 x 8
##   name      birth death lifespan occupation  country  id      dataset
##   <chr>    <int> <int>   <int> <chr>    <chr>  <chr>  <chr>
## 1 HÃ|dde      600   705    105 Catholic pr~ United K~ http://www.w~ wikida~
## 2 Birinus      600   650     50 priest      France  http://www.w~ wikida~
## 3 Angelomus o~  780   855     75 monk      France  http://www.w~ wikida~
## 4 Aymard of C~  910   965     55 <NA>      France  http://www.w~ wikida~
## 5 Wichburg     960  1030     70 <NA>      France  http://www.w~ wikida~
## 6 Geoffrey II~ 1000  1046     46 aristocrat  France  http://www.w~ wikida~
```

Exporting Data to .csv

```
write.csv(df,"wikidata_processed.csv", row.names = FALSE)
```

Data Quality and Exploration

```
dim(df)
```

```
## [1] 2921    8
```

```
range(df$birth)
```

```
## [1] -800 2010
```

```
range(df$death)
```

```
## [1] -760 2038
```

```
range(df$lifespan)
```

```
## [1] 0 200
```

There are 2,921 persons in our European persons data set. The minimum and maximum values for the birth year column are -800 and 2010, and -760 and 2038 for the death year column. The quality of the data is questionable, as there is a person with a death year of 2038, which is not in the past. In addition, I display the entries with a lifespan of 0 below. The birth and death years for most of them seem too conveniently located at the start of each century, and their birth and death year information is questionable as the occupations indicate they should at least be adults or children, not persons of lifespan of 0 years.

```
filter(df, lifespan==0)
```

```
## # A tibble: 17 x 8
##   name      birth death lifespan occupation  country id          dataset
##   <chr>    <int> <int>    <int> <chr>    <chr> <chr>    <chr>
## 1 Amaro      1300  1300      0 pilgrim  France http://www.wi~ wikida~
## 2 Julien Bor~ 1700  1700      0 chronicler France http://www.wi~ wikida~
## 3 Q16719798  1800  1800      0 opera singer France http://www.wi~ wikida~
## 4 William Ga~ 1801  1801      0 lithographer France http://www.wi~ wikida~
## 5 Aline Marn~ 2000  2000      0 film actor  France http://www.wi~ wikida~
## 6 Yitzhak Sa~ 1500  1500      0 <NA>       Germany http://www.wi~ wikida~
## 7 Paul of Na~ 300   300      0 priest      Italy  http://www.wi~ wikida~
## 8 Austromoine 300   300      0 Catholic pr~ Italy  http://www.wi~ wikida~
## 9 Gaius Juli~ 400   400      0 writer      Italy  http://www.wi~ wikida~
## 10 Kate Serje~ 1918  1918      0 actor       United ~ http://www.wi~ wikida~
## 11 Juan de Es~ 1550  1550      0 painter     Spain  http://www.wi~ wikida~
## 12 Isabel Rod~ 1600  1600      0 physician   Spain  http://www.wi~ wikida~
## 13 RamÃ³n PÃ©~ 2000  2000      0 writer      Spain  http://www.wi~ wikida~
## 14 Saeivius Ni~ -250  -250      0 grammarian  Greece http://www.wi~ wikida~
## 15 Prosdocius 100   100      0 Catholic pr~ Greece http://www.wi~ wikida~
## 16 Lopo de Al~ 1400  1400      0 politician  Portugal http://www.wi~ wikida~
## 17 JerÃ³nimo ~ 1900  1900      0 illustrator Portugal http://www.wi~ wikida~
```

There are 421 different occupations in the dataset. The most common ones are painter, actor, Catholic priest, military personnel, and politician. The possibility of overlap between the occupation labels, or synonymous terms, have not been explored here (eg. I noticed the labels “priest” and “Catholic priest” in the data)

```
length(unique(df$occupation))
```

```
## [1] 421
```

```
df %>%
  filter(!is.na(occupation)) %>%
  group_by(occupation) %>%
  summarise(count = n()) %>%
  arrange(-count) %>%
  top_n(n = 5, wt = count)
```

```
## # A tibble: 5 x 2
##   occupation      count
##   <chr>          <int>
## 1 painter         161
## 2 actor           160
## 3 Catholic priest  146
## 4 military personnel 146
## 5 politician      144
```

The majority of the persons come from (more specifically, were born in) Germany, France, and the UK.

```
df %>%
  group_by(country) %>%
  summarise(count = n()) %>%
  arrange(-count) %>%
  top_n(n = 8, wt = count)
```

```
## # A tibble: 8 x 2
##   country      count
##   <chr>        <int>
## 1 Germany       712
## 2 France        614
## 3 United Kingdom 438
## 4 Italy         377
## 5 Spain         285
## 6 Russia        261
## 7 Portugal      143
## 8 Greece        90
```

Example Plot

A scatter plot of the data, with a moving average line in blue. We limit the x-axis to [1500,2022] as most of the data is concentrated in that timeframe. We limit the upper bound to 120, as lifespans above 120 may seem nonsensical.

A horizontal strip of deaths can be noticed around the time of World War 2. The European longevity trend appears to be relatively flat around 55-60 years from 1500 to 1950. After 1950, the average longevity increases to a peak of around 75, but appears to flatten around the new millenia.

```
plot <-
  ggplot(data=df, aes(x=death, y=lifespan)) +
  geom_point(alpha=0.2) +
  ggtitle("European Lifespans") +
  theme_bw() +
  ylab("Lifespan") +
  scale_x_continuous(breaks = seq(-800, max(df$death), 200), name = "Year of Death")
```

```
plot +
  scale_x_continuous(breaks = seq(1000, 2000, 100), limits=c(1500,2022), name = "Year of Death")
+
  scale_y_continuous(breaks = seq(0, 120, 10), limits=c(0,120)) +
  geom_smooth(method="loess", se = FALSE) +
  labs(caption = "Lifespan [0, 120]\nYear of Death [1500, 2000]")
```

