Variables-Avance de Trabajo Grupal

Links de las Fuentes de Datos

<https://www.worldatlas.com/articles/the-highest-literacy-rates-in-the-world.html> Literacy by worldatlas

<http://hdr.undp.org/en/content/education-index>  Schooling

<https://population.un.org/wup/Download/> Population rate by UN

<http://worldpopulationreview.com/countries/education-rankings-by-country/#dataTable> PISA TEST

<https://www.cia.gov/library/publications/resources/the-world-factbook/fields/370.html> Literacy by CIA

Códigos y trabajos en R Studio Cloud:

Primero limpiaré la información de la tasa de alfabetización otorgado por la CIA

```{r}

library(htmltab)

literacy=htmltab(doc="https://www.cia.gov/library/publications/resources/the-world-factbook/fields/370.html", which='//\*[@id="fieldListing"]')

```

```{r}

str(literacy)

```

```{r}

library(stringr)

str\_extract\_all(literacy$Literacy,pattern="(\\-\*\\d+\\.\*\\d\*)(?=\\%)",simplify = T)

PATRON="(\\-\*\\d+\\.\*\\d\*)(?=\\%)"

COLSUCIA=literacy$Literacy

# UNA COLUMNA

literacy$pop\_lit=str\_extract\_all(string = COLSUCIA,pattern=PATRON,simplify = T)[,1]

# OTRA COLUMNA

literacy$male\_lit=str\_extract\_all(string = COLSUCIA,pattern=PATRON,simplify = T)[,2]

literacy$female\_lit=str\_extract\_all(string = COLSUCIA,pattern=PATRON,simplify = T)[,3]

```

```{r}

head(literacy)

```

```{r}

literacy$Literacy=NULL

```

```{r}

literacy[,-1]=lapply(literacy[,-1], as.numeric)

```

```{r}

str(literacy)

```

```{r}

literacy[!complete.cases(literacy),]

```

```{r}

literacy$Country=trimws(literacy$Country,whitespace = "[\\h\\v]")

```

```{r}

names(literacy)=c("Country", "Literacy Rate", "Male Literacy Rate", "Female Literacy Rate")

```

```{r}

head(literacy)

```

Como extra, importaré y limpiaré otra tabla de alfabetización del worldatlas ya que esta posee más paises en su muestra

```{r}

library(htmltab)

literacy2=htmltab(doc="https://www.worldatlas.com/articles/the-highest-literacy-rates-in-the-world.html", which='//\*[@id="artReg-table"]/table')

```

```{r}

literacy2$Rank=NULL

```

```{r}

names(literacy2)=c("Country", "Literacy Rate")

```

```{r}

literacy2=literacy2[-c(196),]

```

Las siguientes variables son las usaré para la calidad de la educación. La primera sería la prueba PISA

```{r}

lkCSV="https://github.com/Geanze/TRABAJO-GRUPAL-ESTAD2/raw/master/Pisa%20rank.csv"

PISA=import(lkCSV)

```

```{r}

dim(PISA)

```

```{r}

PISA$cca2=NULL

PISA$pop2019=NULL

PISA$mathScore=NULL

PISA$readingScore=NULL

PISA$scienceScore=NULL

```

```{r}

names(PISA)=c("Country", "PISA SCORE")

```

```{r}

str(PISA)

```

La segunda variable que tiene relación con la calidad de educación es la Tasa de Escolaridad.

```{r}

lkXLSX="https://github.com/Geanze/TRABAJO-GRUPAL-ESTAD2/raw/master/education-index.xlsx"

Schooling=import(lkXLSX)

```

```{r}

str(Schooling)

```

```{r}

Schooling$`HDI Rank`=NULL

Schooling$`1980`=NULL

Schooling$`1985`=NULL

Schooling$`1990`=NULL

Schooling$`1995`=NULL

Schooling$`2000`=NULL

Schooling$`2005`=NULL

Schooling$`2006`=NULL

Schooling$`2007`=NULL

Schooling$`2008`=NULL

Schooling$`2009`=NULL

Schooling$`2010`=NULL

Schooling$`2011`=NULL

Schooling$`2012`=NULL

```

```{r}

names(Schooling)=c("Country", "Schooling Rate")

```

```{r}

Schooling[c(2)]= lapply(Schooling[c(2)], as.numeric)

```

```{r}

Schooling[complete.cases(Schooling),]

Schooling=na.omit(Schooling)

```

La última variable es la Tasa de Urbanización.

```{r}

lkXLS="https://github.com/Geanze/TRABAJO-GRUPAL-ESTAD2/raw/master/WUP2018-F01-Total\_Urban\_Rural.xls"

urban=import(lkXLS)

```

```{r}

str(urban)

```

```{r}

urban$Urban=NULL

urban$Rural=NULL

urban$Total=NULL

```

```{r}

names(urban)=c("Country", "Urban Rate")

```

```{r}

urban[complete.cases(urban),]

urban[c(1)]=lapply(urban[c(1)], trimws,whitespace = "[\\h\\v]")

```