# Statistical Analysis of the EU Data in R

**Cumali Bereket s211060 BDA SGGW**

Project\_Cumali\_Bereket

The inactivity rate is a measure of how many people are not in the workforce. It has a big impact on the socio-economic development of a country or region, because it affects how much productive work is done. Countries and regions can try to reduce the inactivity rate by doing things like encouraging people to get back into the workforce.

# Importing the data

After searching for data on eurostat, we imported it using the code “lfst\_r\_lfp2actrtn”.

library(eurostat)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
library(onewaytests)  
library(ggpubr)

## Loading required package: ggplot2

library(FSA)

## Registered S3 methods overwritten by 'FSA':  
## method from  
## confint.boot car   
## hist.boot car

## ## FSA v0.9.3. See citation('FSA') if used in publication.  
## ## Run fishR() for related website and fishR('IFAR') for related book.

library(ggplot2)  
library(tidyverse)

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ tibble 3.1.8 ✔ stringr 1.5.0  
## ✔ readr 2.1.3 ✔ forcats 0.5.2  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(stringr)  
library(sf)

## Linking to GEOS 3.9.3, GDAL 3.5.2, PROJ 8.2.1; sf\_use\_s2() is TRUE

library(ggpubr)  
library(onewaytests)

data <- get\_eurostat("lfst\_r\_lfp2actrtn", time\_format = "num")

I deal with the data : Female inactivity rate aged 20-64.And I am starting by examining how the data set is constructed.

str(data)

## tibble [3,702,136 × 8] (S3: tbl\_df/tbl/data.frame)  
## $ citizen: chr [1:3702136] "EU27\_2020\_FOR" "EU27\_2020\_FOR" "EU27\_2020\_FOR" "EU27\_2020\_FOR" ...  
## $ isced11: chr [1:3702136] "ED0-2" "ED0-2" "ED0-2" "ED0-2" ...  
## $ age : chr [1:3702136] "Y15-64" "Y15-64" "Y15-64" "Y15-64" ...  
## $ sex : chr [1:3702136] "F" "F" "F" "F" ...  
## $ unit : chr [1:3702136] "PC" "PC" "PC" "PC" ...  
## $ geo : chr [1:3702136] "AT" "AT1" "AT11" "AT12" ...  
## $ time : num [1:3702136] 2021 2021 2021 2021 2021 ...  
## $ values : num [1:3702136] 56.5 55 NA 59.6 53.3 51.3 NA 53.4 61.7 57.9 ...

data$citizen <- as.factor(data$citizen)  
data$isced11 <- as.factor(data$isced11)  
data$age <- as.factor(data$age)  
data$sex <- as.factor(data$sex)  
data$unit <- as.factor(data$unit)  
data$geo <- as.factor(data$geo)  
data$time <- as.factor(data$time)

str(data)

## tibble [3,702,136 × 8] (S3: tbl\_df/tbl/data.frame)  
## $ citizen: Factor w/ 11 levels "EU15\_FOR","EU27\_2020\_FOR",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ isced11: Factor w/ 5 levels "ED0-2","ED3\_4",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ age : Factor w/ 4 levels "Y15-64","Y20-64",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ sex : Factor w/ 3 levels "F","M","T": 1 1 1 1 1 1 1 1 1 1 ...  
## $ unit : Factor w/ 1 level "PC": 1 1 1 1 1 1 1 1 1 1 ...  
## $ geo : Factor w/ 506 levels "AT","AT1","AT11",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ time : Factor w/ 23 levels "1999","2000",..: 23 23 23 23 23 23 23 23 23 23 ...  
## $ values : num [1:3702136] 56.5 55 NA 59.6 53.3 51.3 NA 53.4 61.7 57.9 ...

With the exception of geo, where there are too many values, we are able to examine the levels of each variable and factor.

levels(data$citizen)

## [1] "EU15\_FOR" "EU27\_2020\_FOR" "EU28\_FOR" "FOR"   
## [5] "NAT" "NEU15\_FOR" "NEU27\_2020\_FOR" "NEU28\_FOR"   
## [9] "NRP" "STLS" "TOTAL"

levels(data$isced11)

## [1] "ED0-2" "ED3\_4" "ED5-8" "NRP" "TOTAL"

levels(data$age)

## [1] "Y15-64" "Y20-64" "Y25-54" "Y55-64"

levels(data$sex)

## [1] "F" "M" "T"

mydata = data %>%  
 filter(sex=="F" & age=="Y20-64" & (time=="2019" | time=="2020" | time=="2021") & citizen=="TOTAL" & !grepl("UK",geo,fixed =FALSE) & !grepl("ZZ",geo,fixed =FALSE) & geo!="FRY1" & geo!="FRY2" & geo!="FRY3" & geo!="FRY4" & geo!="FRY5" ) %>%  
 mutate(Inactivity\_rate = 100-values)

mydata <- validate\_nuts\_regions(mydata, geo\_var = "geo", nuts\_year = 2021)

Now let’s examine the various levels (country is nut 0) and determine if they are all valid:

table(mydata$typology)

##   
## country nuts\_level\_1 nuts\_level\_2   
## 429 1399 3520

table(mydata$valid\_2021)

##   
## FALSE TRUE   
## 84 5348

mydata$citizen <- as.factor(mydata$citizen)  
mydata$isced11 <- as.factor(mydata$isced11)  
mydata$age <- as.factor(mydata$age)  
mydata$sex <- as.factor(mydata$sex)  
mydata$unit <- as.factor(mydata$unit)  
mydata$geo <- as.factor(mydata$geo)  
mydata$time <- as.factor(mydata$time)

mydata$isced11<-as.character(mydata$isced11)  
mydata <- mydata %>%   
 mutate( edu =   
 case\_when(  
 isced11 == "ED0-2" ~ "edu\_low",  
 isced11 == "ED3\_4" ~ "edu\_medium",  
 isced11 == "ED5-8" ~ "edu\_high",  
 isced11 == "NRP" ~ "NRP",  
 isced11 == "TOTAL" ~ "edu\_total"  
 )  
 ) %>%  
 select(-isced11)  
mydata$edu<-as.factor(mydata$edu)

\*We just save the countries for Nuts 0.

mydata\_nuts0 = mydata %>%  
 filter(typology=="country")

We take out the countries for Nuts 2 and only keep the regions.

mydata\_nuts2 = mydata %>%  
 filter(typology=="nuts\_level\_2" )

Regarding the differences in the number of lines, same as with Nuts 0.

First, we compile a list of our variables.

summary(mydata\_nuts0)

## citizen age sex unit geo time   
## TOTAL:429 Y20-64:429 F:429 PC:429 CH : 15 2019:147   
## CZ : 15 2020:147   
## DE : 15 2021:135   
## DK : 15   
## FR : 15   
## IE : 15   
## (Other):339   
## values Inactivity\_rate typology valid\_2021   
## Min. :25.40 Min. : 5.10 Length:429 Mode:logical   
## 1st Qu.:59.98 1st Qu.:18.00 Class :character TRUE:429   
## Median :73.10 Median :26.90 Mode :character   
## Mean :70.10 Mean :29.90   
## 3rd Qu.:82.00 3rd Qu.:40.02   
## Max. :94.90 Max. :74.60   
## NA's :9 NA's :9   
## edu   
## edu\_high :99   
## edu\_low :99   
## edu\_medium:99   
## edu\_total :99   
## NRP :33   
##   
##

Out of 429 values, we can see that 9 are missing for the inactivity rate. Let’s take a closer look at them:

na\_values = mydata\_nuts0 %>% filter(is.na(Inactivity\_rate))

summary(na\_values)

## citizen age sex unit geo time values   
## TOTAL:9 Y20-64:9 F:9 PC:9 CZ :3 2019:3 Min. : NA   
## LV :2 2020:2 1st Qu.: NA   
## DE :1 2021:4 Median : NA   
## FI :1 Mean :NaN   
## IS :1 3rd Qu.: NA   
## NO :1 Max. : NA   
## (Other):0 NA's :9   
## Inactivity\_rate typology valid\_2021 edu   
## Min. : NA Length:9 Mode:logical edu\_high :0   
## 1st Qu.: NA Class :character TRUE:9 edu\_low :0   
## Median : NA Mode :character edu\_medium:0   
## Mean :NaN edu\_total :0   
## 3rd Qu.: NA NRP :9   
## Max. : NA   
## NA's :9

We quickly realize that the NRP education level contains all of the missing data. We are going to remove this modality from the remainder of the analysis because we only have nine values for it compared to 100 for the other modalities of this variable.

mydata\_nuts0<-mydata\_nuts0 %>%  
 filter(edu != "NRP")

min(mydata\_nuts0$Inactivity\_rate)

## [1] 6.9

max(mydata\_nuts0$Inactivity\_rate)

## [1] 74.6

For the variable of the inactivity rate, we do not observe any outliers.

table(droplevels(mydata\_nuts0$geo))

##   
## AT BE BG CH CY CZ DE DK EE EL ES FI FR HR HU IE IS IT LT LU LV ME MK MT NL NO   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 8 8 12 12 12   
## PL PT RO RS SE SI SK TR   
## 12 12 12 12 12 12 12 8

3 countries have only 8 values because they arrived or left the European Union in these 3 years. We start by making a summary of our data according to the 4 levels of education:

inactivity2 <- mydata\_nuts0 %>%   
 pivot\_wider(names\_from = edu, values\_from = Inactivity\_rate)

summary(inactivity2)[,-c(seq(1,9,1))]

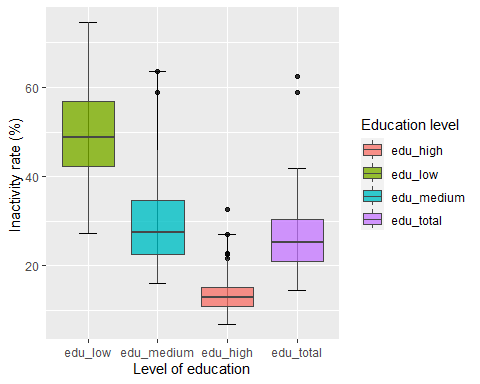
## edu\_low edu\_medium edu\_high edu\_total   
## Min. :27.20 Min. :16.10 Min. : 6.90 Min. :14.60   
## 1st Qu.:42.20 1st Qu.:22.50 1st Qu.:10.90 1st Qu.:21.10   
## Median :48.70 Median :27.40 Median :12.90 Median :25.30   
## Mean :49.48 Mean :28.77 Mean :13.55 Mean :26.76   
## 3rd Qu.:56.90 3rd Qu.:34.60 3rd Qu.:15.15 3rd Qu.:30.35   
## Max. :74.60 Max. :63.60 Max. :32.70 Max. :62.50   
## NA's :296 NA's :296 NA's :296 NA's :296

In comparison to the other categories, the edu\_high category has a significantly lower average inactivity rate. On the other hand, the inactivity rate is significantly higher in the edu\_low category.

level\_orderX <- c("edu\_low", "edu\_medium","edu\_high", "edu\_total")

ggplot(mydata\_nuts0, aes(x = factor(edu, level=level\_orderX), y = Inactivity\_rate, fill=edu)) +   
 xlab("Level of education") + ylab("Inactivity rate (%) ")+  
 stat\_boxplot(na.rm = TRUE,  
 coef = 3, #default coef = 1.5  
 geom = "errorbar",  
 order = c("edu\_low", "edu\_medium", "edu\_high"),  
 width = 0.25) +   
 geom\_boxplot(alpha = 0.8, # Fill transparency  
 colour = "#474747", # Border color  
 na.rm = TRUE,  
 outlier.colour = 1) +  
guides(fill = guide\_legend(title = "Education level"))

## Warning in stat\_boxplot(na.rm = TRUE, coef = 3, geom = "errorbar", order =  
## c("edu\_low", : Ignoring unknown parameters: `order`



The boxplots above show what we were talking about before: the edu\_high category has low inactivity, edu\_low has high inactivity, edu\_medium and edu\_total are about the same level.

First of all we make a summary of our variables.

summary(mydata\_nuts2)

## citizen age sex unit geo time   
## TOTAL:3520 Y20-64:3520 F:3520 PC:3520 CH01 : 15 2019:1204   
## CH02 : 15 2020:1209   
## CH03 : 15 2021:1107   
## CH04 : 15   
## CH05 : 15   
## CH06 : 15   
## (Other):3430   
## values Inactivity\_rate typology valid\_2021   
## Min. :14.5 Min. : 5.2 Length:3520 Mode:logical   
## 1st Qu.:59.4 1st Qu.:18.3 Class :character TRUE:3520   
## Median :72.5 Median :27.5 Mode :character   
## Mean :69.0 Mean :31.0   
## 3rd Qu.:81.7 3rd Qu.:40.6   
## Max. :94.8 Max. :85.5   
## NA's :252 NA's :252   
## edu   
## edu\_high :809   
## edu\_low :809   
## edu\_medium:809   
## edu\_total :809   
## NRP :284   
##   
##

On 3520 values, we can see that there are 252 missing values for the inactivity rate. Let’s look at them in detail:

na\_values = mydata\_nuts2 %>% filter(is.na(Inactivity\_rate))

summary(na\_values)

## citizen age sex unit geo time   
## TOTAL:252 Y20-64:252 F:252 PC:252 DE50 : 4 2019:74   
## DE80 : 4 2020:97   
## DEB2 : 4 2021:81   
## FI20 : 4   
## CH07 : 3   
## CZ02 : 3   
## (Other):230   
## values Inactivity\_rate typology valid\_2021   
## Min. : NA Min. : NA Length:252 Mode:logical   
## 1st Qu.: NA 1st Qu.: NA Class :character TRUE:252   
## Median : NA Median : NA Mode :character   
## Mean :NaN Mean :NaN   
## 3rd Qu.: NA 3rd Qu.: NA   
## Max. : NA Max. : NA   
## NA's :252 NA's :252   
## edu   
## edu\_high : 0   
## edu\_low : 25   
## edu\_medium: 0   
## edu\_total : 0   
## NRP :227   
##   
##

Contrary to before, missing values are more distributed here, whether it is for the education level, the region or even the year.

As with Nuts 0, we will remove NRP in a first step

mydata\_nuts2<-mydata\_nuts2 %>%  
 filter(edu != "NRP")

min(mydata\_nuts2$Inactivity\_rate)

## [1] NA

max(mydata\_nuts2$Inactivity\_rate)

## [1] NA

When we examine our data, , we have a single case where the inactivity rate is equal to 0 for the ES63 region in 2020 and for the high education category. We decide to keep the value anyway.

table(droplevels(mydata\_nuts2$geo))

##   
## AT11 AT12 AT13 AT21 AT22 AT31 AT32 AT33 AT34 BE10 BE21 BE22 BE23 BE24 BE25 BE31   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## BE32 BE33 BE34 BE35 BG31 BG32 BG33 BG34 BG41 BG42 CH01 CH02 CH03 CH04 CH05 CH06   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## CH07 CY00 CZ01 CZ02 CZ03 CZ04 CZ05 CZ06 CZ07 CZ08 DE11 DE12 DE13 DE14 DE21 DE22   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## DE23 DE24 DE25 DE26 DE27 DE30 DE40 DE50 DE60 DE71 DE72 DE73 DE80 DE91 DE92 DE93   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## DE94 DEA1 DEA2 DEA3 DEA4 DEA5 DEB1 DEB2 DEB3 DEC0 DED2 DED4 DED5 DEE0 DEF0 DEG0   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## DK01 DK02 DK03 DK04 DK05 EE00 EL30 EL41 EL42 EL43 EL51 EL52 EL53 EL54 EL61 EL62   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## EL63 EL64 EL65 ES11 ES12 ES13 ES21 ES22 ES23 ES24 ES30 ES41 ES42 ES43 ES51 ES52   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## ES53 ES61 ES62 ES63 ES64 ES70 FI19 FI1B FI1C FI1D FI20 FR10 FRB0 FRC1 FRC2 FRD1   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## FRD2 FRE1 FRE2 FRF1 FRF2 FRF3 FRG0 FRH0 FRI1 FRI2 FRI3 FRJ1 FRJ2 FRK1 FRK2 FRL0   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## FRM0 HR02 HR03 HR05 HR06 HU11 HU12 HU21 HU22 HU23 HU31 HU32 HU33 IE04 IE05 IE06   
## 12 4 12 4 4 12 12 12 12 12 12 12 12 12 12 12   
## IS00 ITC1 ITC2 ITC3 ITC4 ITF1 ITF2 ITF3 ITF4 ITF5 ITF6 ITG1 ITG2 ITH1 ITH2 ITH3   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## ITH4 ITH5 ITI1 ITI2 ITI3 ITI4 LT01 LT02 LU00 LV00 ME00 MK00 MT00 NL11 NL12 NL13   
## 12 12 12 12 12 12 12 12 12 12 8 8 12 12 12 12   
## NL21 NL22 NL23 NL31 NL32 NL33 NL34 NL41 NL42 NO02 NO06 NO07 NO08 NO09 NO0A PL21   
## 12 12 12 12 12 12 12 12 12 12 12 12 4 4 4 12   
## PL22 PL41 PL42 PL43 PL51 PL52 PL61 PL62 PL63 PL71 PL72 PL81 PL82 PL84 PL91 PL92   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## PT11 PT15 PT16 PT17 PT18 PT20 PT30 RO11 RO12 RO21 RO22 RO31 RO32 RO41 RO42 RS11   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## RS12 RS21 RS22 SE11 SE12 SE21 SE22 SE23 SE31 SE32 SE33 SI03 SI04 SK01 SK02 SK03   
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12   
## SK04 TR10 TR21 TR22 TR31 TR32 TR33 TR41 TR42 TR51 TR52 TR61 TR62 TR63 TR71 TR72   
## 12 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8   
## TR81 TR82 TR83 TR90 TRA1 TRA2 TRB1 TRB2 TRC1 TRC2 TRC3   
## 8 8 8 8 8 8 8 8 8 8 8

The same as in the previous case, the regions that are not equal to 12 come from countries that have arrived or left the European Union in these 3 years.

We start by making a summary of our data according to the 4 levels of education:

inactivity2 <- mydata\_nuts2 %>%   
 pivot\_wider(names\_from = edu, values\_from = Inactivity\_rate)

summary(inactivity2)[,-c(seq(1,9,1))]

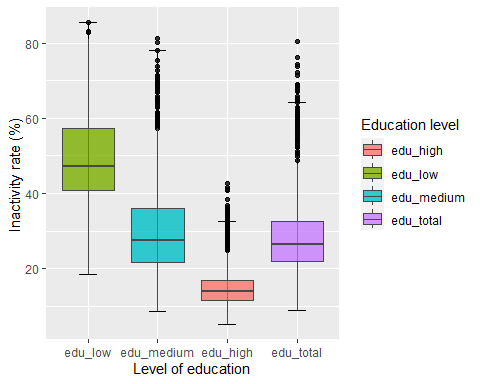
## edu\_low edu\_medium edu\_high edu\_total   
## Min. :18.4 Min. : 8.70 Min. : 5.20 Min. : 9.00   
## 1st Qu.:40.7 1st Qu.:21.70 1st Qu.:11.50 1st Qu.:22.00   
## Median :47.2 Median :27.50 Median :14.00 Median :26.40   
## Mean :49.3 Mean :30.54 Mean :15.34 Mean :29.22   
## 3rd Qu.:57.2 3rd Qu.:35.90 3rd Qu.:16.80 3rd Qu.:32.60   
## Max. :85.5 Max. :81.20 Max. :42.60 Max. :80.40   
## NA's :2444 NA's :2419 NA's :2419 NA's :2419

In comparison to the other categories, the edu\_high category has a significantly lower average inactivity rate. On the other hand, the inactivity rate is significantly higher in the edu\_low category. Similar to NUTS 0, exactly.

level\_orderX <- c("edu\_low", "edu\_medium","edu\_high", "edu\_total")

ggplot(mydata\_nuts2, aes(x = factor(edu, level=level\_orderX), y = Inactivity\_rate, fill=edu)) +   
 xlab("Level of education") + ylab("Inactivity rate (%) ")+  
 stat\_boxplot(na.rm = TRUE,  
 coef = 3, #default coef = 1.5  
 geom = "errorbar",  
 order = c("edu\_low", "edu\_medium", "edu\_high"),  
 width = 0.25) +   
 geom\_boxplot(alpha = 0.8, # Fill transparency  
 colour = "#474747", # Border color  
 na.rm = TRUE,  
 outlier.colour = 1) +  
guides(fill = guide\_legend(title = "Education level"))

## Warning in stat\_boxplot(na.rm = TRUE, coef = 3, geom = "errorbar", order =  
## c("edu\_low", : Ignoring unknown parameters: `order`



The boxplots above show what we were talking about before: the edu\_high category has low inactivity, edu\_low has high inactivity and edu\_medium and edu\_total are about the same level. Same as with Nuts 0.

map\_eu <- get\_eurostat\_geospatial(  
 output\_class = "sf",  
 resolution = "60",  
 nuts\_level = 0,  
 year = "2021",  
 cache = TRUE,  
 update\_cache = FALSE,  
 cache\_dir = NULL,  
 crs = "4326",  
 make\_valid = FALSE  
)

## Object cached at C:\Users\CUMALI~1\AppData\Local\Temp\RtmpsFpIVw/eurostat/sf60020214326.RData

## Reading cache file C:\Users\CUMALI~1\AppData\Local\Temp\RtmpsFpIVw/eurostat/sf60020214326.RData

## sf at resolution 1: 60 from year 2021 read from cache file: C:\Users\CUMALI~1\AppData\Local\Temp\RtmpsFpIVw/eurostat/sf60020214326.RData

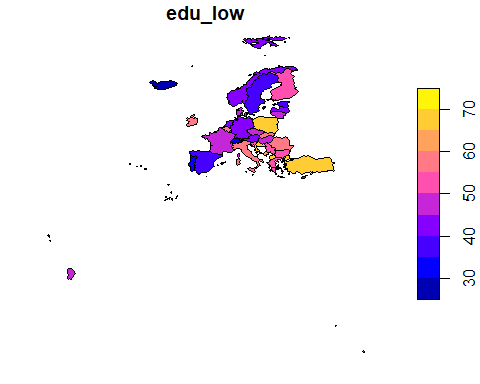
## Warning in get\_eurostat\_geospatial(output\_class = "sf", resolution = "60", :  
## Default of 'make\_valid' for 'output\_class="sf"' will be changed in the future  
## (see function details).

map0 <- mydata\_nuts0 %>%   
 pivot\_wider(names\_from = edu, values\_from = Inactivity\_rate)

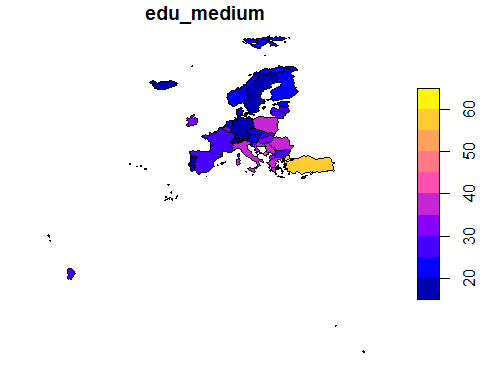
map\_data0 <- inner\_join(map\_eu, map0)

## Joining, by = "geo"

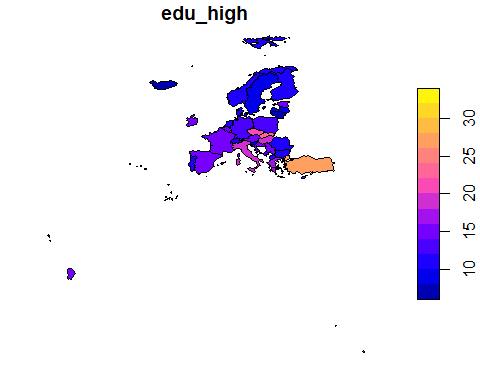
map0\_low <- map\_data0 %>%  
 select(geometry, edu\_low)  
  
plot(map0\_low)



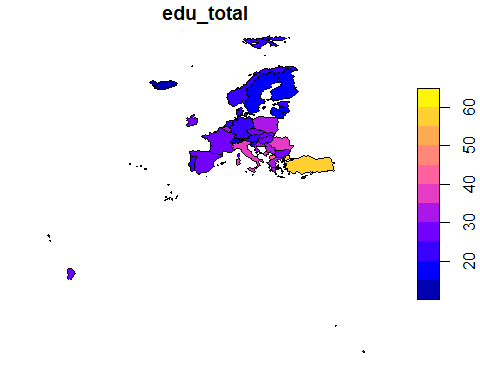
map0\_medium <- map\_data0 %>%  
 select(geometry, edu\_medium)  
  
plot(map0\_medium)



map0\_high <- map\_data0 %>%  
 select(geometry, edu\_high)  
  
plot(map0\_high)



map0\_total <- map\_data0 %>%  
 select(geometry, edu\_total)  
  
plot(map0\_total)



map\_eu2 <- get\_eurostat\_geospatial(  
 output\_class = "sf",  
 resolution = "60",  
 nuts\_level = 2,  
 year = "2021",  
 cache = TRUE,  
 update\_cache = FALSE,  
 cache\_dir = NULL,  
 crs = "4326",  
 make\_valid = FALSE  
)

## Object cached at C:\Users\CUMALI~1\AppData\Local\Temp\RtmpsFpIVw/eurostat/sf60220214326.RData

## Reading cache file C:\Users\CUMALI~1\AppData\Local\Temp\RtmpsFpIVw/eurostat/sf60220214326.RData

## sf at resolution 1: 60 from year 2021 read from cache file: C:\Users\CUMALI~1\AppData\Local\Temp\RtmpsFpIVw/eurostat/sf60220214326.RData

## Warning in get\_eurostat\_geospatial(output\_class = "sf", resolution = "60", :  
## Default of 'make\_valid' for 'output\_class="sf"' will be changed in the future  
## (see function details).

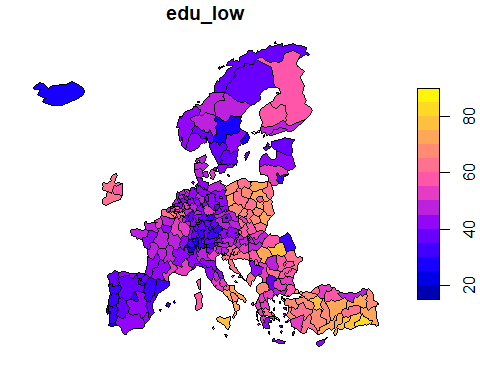
map2 <- mydata\_nuts2 %>%   
 pivot\_wider(names\_from = edu, values\_from = Inactivity\_rate)

map\_data2 <- inner\_join(map\_eu2, map2)

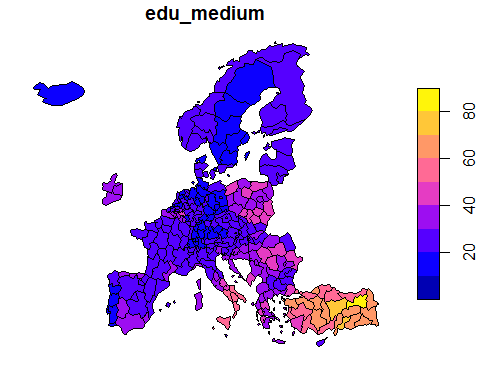
## Joining, by = "geo"

map\_data\_cont2 <- map\_data2 %>%  
 # Exclude Islands from analysis  
 filter(!NUTS\_ID %in% c(  
 "ES63", #ceuta  
 "ES64", #melilla  
 "ES70", #Canarias  
 "PT30", #PT Madeira  
 "PT20" , #PR Acores  
 "FRY1",  
 "FRY2" ,  
 "FRY3" ,  
 "FRY4" ,  
 "FRY5"  
 )  
 )

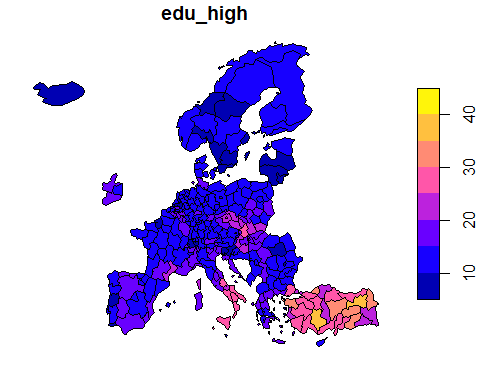
map2\_low <- map\_data\_cont2 %>%  
 select(geometry, edu\_low)  
  
plot(map2\_low)



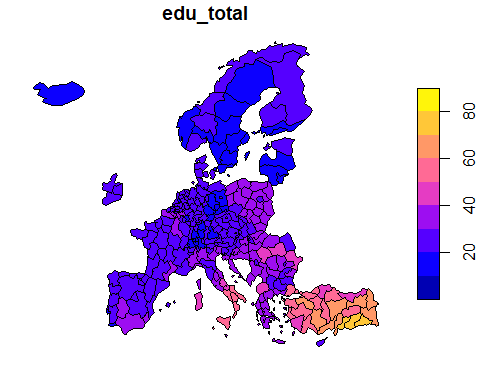
map2\_medium <- map\_data\_cont2 %>%  
 select(geometry, edu\_medium)  
  
plot(map2\_medium)



map2\_high <- map\_data\_cont2 %>%  
 select(geometry, edu\_high)  
  
plot(map2\_high)



map2\_total <- map\_data\_cont2 %>%  
 select(geometry, edu\_total)  
  
plot(map2\_total)



We begin by categorizing education levels. We only need to get rid of the mode “edu total” to accomplish this.

mydata\_nuts0a <- mydata\_nuts0 %>%  
 filter (edu!="edu\_total")

Then, for each year, we divide our dataset in 3.

mydata\_nuts0a2019 <- mydata\_nuts0a %>%  
 filter (time=="2019")  
  
mydata\_nuts0a2020 <- mydata\_nuts0a %>%  
 filter (time=="2020")  
  
mydata\_nuts0a2021 <- mydata\_nuts0a %>%  
 filter (time=="2021")

mydata\_nuts0a2019\_low <- mydata\_nuts0a2019 %>%  
 filter (edu=="edu\_low")  
mydata\_nuts0a2019\_medium <- mydata\_nuts0a2019 %>%  
 filter (edu=="edu\_medium")  
mydata\_nuts0a2019\_high <- mydata\_nuts0a2019 %>%  
 filter (edu=="edu\_high")

We test the normality of the distribution with the Shapiro test with the following hypothesis

H0 : The data come from a normal distribution H1 : The data do not come from a normal distribution

shapiro.test(mydata\_nuts0a2019\_low$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0a2019\_low$Inactivity\_rate  
## W = 0.98474, p-value = 0.9053

shapiro.test(mydata\_nuts0a2019\_medium$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0a2019\_medium$Inactivity\_rate  
## W = 0.91742, p-value = 0.01367

shapiro.test(mydata\_nuts0a2019\_high$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0a2019\_high$Inactivity\_rate  
## W = 0.93384, p-value = 0.04051

model <- lm(Inactivity\_rate ~ edu, data = mydata\_nuts0a2019)  
shapiro.test(residuals(model))

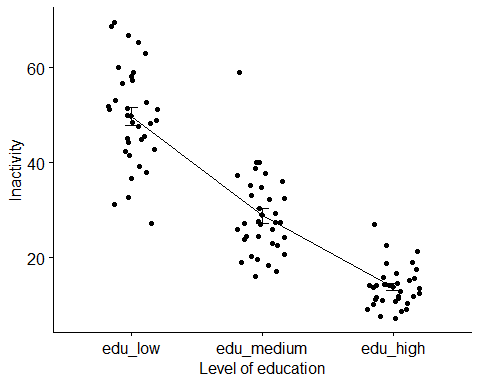
##   
## Shapiro-Wilk normality test  
##   
## data: residuals(model)  
## W = 0.97313, p-value = 0.03541

The Shapiro test confirms this with a pvalue < alpha(5%).

We cannot say that the sample follows a normal distribution, so we will therefore use a non-parametric Anova.

We can visualize our data.

data\_anova\_nuts0a2019 <- mydata\_nuts0a2019 %>%  
 select(geo, edu, Inactivity\_rate)  
  
ggline(data\_anova\_nuts0a2019, x = "edu", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("edu\_low", "edu\_medium", "edu\_high"),  
 ylab = "Inactivity", xlab = "Level of education")



out <- ag.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts0a2019)

##   
## Alexander-Govern Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and edu   
##   
## statistic : 133.1989   
## parameter : 2   
## p.value : 1.191877e-29   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

out2 <- bf.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts0a2019)

##   
## Brown-Forsythe Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and edu   
##   
## statistic : 160.5715   
## num df : 2   
## denom df : 75.41473   
## p.value : 6.58621e-28   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

one.way <- aov(Inactivity\_rate ~ edu, data = data\_anova\_nuts0a2019)  
summary(one.way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## edu 2 22035 11017 160.6 <2e-16 \*\*\*  
## Residuals 99 6793 69   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

kruskal.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts0a2019)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Inactivity\_rate by edu  
## Kruskal-Wallis chi-squared = 81.287, df = 2, p-value < 2.2e-16

mydata\_nuts0a2020\_low <- mydata\_nuts0a2020 %>%  
 filter (edu=="edu\_low")  
mydata\_nuts0a2020\_medium <- mydata\_nuts0a2020 %>%  
 filter (edu=="edu\_medium")  
mydata\_nuts0a2020\_high <- mydata\_nuts0a2020 %>%  
 filter (edu=="edu\_high")

shapiro.test(mydata\_nuts0a2020\_low$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0a2020\_low$Inactivity\_rate  
## W = 0.98262, p-value = 0.8509

shapiro.test(mydata\_nuts0a2020\_medium$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0a2020\_medium$Inactivity\_rate  
## W = 0.90312, p-value = 0.005553

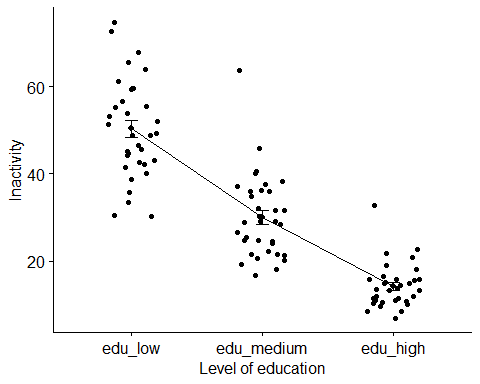
shapiro.test(mydata\_nuts0a2020\_high$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0a2020\_high$Inactivity\_rate  
## W = 0.87249, p-value = 0.0009294

2 of the 3 groups has a pvalue < 0.05, so we reject h0. We cannot say that the sample follows a normal distribution, so we will therefore use a non-parametric Anova.

We can visualize our data.

data\_anova\_nuts0a2020 <- mydata\_nuts0a2020 %>%  
 select(geo, edu, Inactivity\_rate)  
  
ggline(data\_anova\_nuts0a2020, x = "edu", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("edu\_low", "edu\_medium", "edu\_high"),  
 ylab = "Inactivity", xlab = "Level of education")



out <- ag.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts0a2020)

##   
## Alexander-Govern Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and edu   
##   
## statistic : 124.2531   
## parameter : 2   
## p.value : 1.044201e-27   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

out2 <- bf.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts0a2020)

##   
## Brown-Forsythe Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and edu   
##   
## statistic : 138.3873   
## num df : 2   
## denom df : 77.62645   
## p.value : 2.530746e-26   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

one.way <- aov(Inactivity\_rate ~ edu, data = data\_anova\_nuts0a2020)  
summary(one.way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## edu 2 22390 11195 138.4 <2e-16 \*\*\*  
## Residuals 99 8009 81   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

kruskal.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts0a2020)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Inactivity\_rate by edu  
## Kruskal-Wallis chi-squared = 79.248, df = 2, p-value < 2.2e-16

mydata\_nuts0a2021\_low <- mydata\_nuts0a2021 %>%  
 filter (edu=="edu\_low")  
mydata\_nuts0a2021\_medium <- mydata\_nuts0a2021 %>%  
 filter (edu=="edu\_medium")  
mydata\_nuts0a2021\_high <- mydata\_nuts0a2021 %>%  
 filter (edu=="edu\_high")

shapiro.test(mydata\_nuts0a2021\_low$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0a2021\_low$Inactivity\_rate  
## W = 0.96667, p-value = 0.4325

shapiro.test(mydata\_nuts0a2021\_medium$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0a2021\_medium$Inactivity\_rate  
## W = 0.94516, p-value = 0.1147

shapiro.test(mydata\_nuts0a2021\_high$Inactivity\_rate)

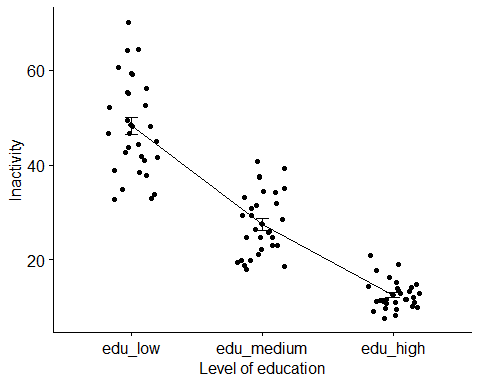
##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0a2021\_high$Inactivity\_rate  
## W = 0.9342, p-value = 0.05711

The 3 groups have a pvalue > 0.05, so we cannot reject h0.

We can say that the sample follows a normal distribution, so we will therefore use a parametric Anova.

We can visualize our data.

data\_anova\_nuts0a2021 <- mydata\_nuts0a2021 %>%  
 select(geo, edu, Inactivity\_rate)  
  
ggline(data\_anova\_nuts0a2021, x = "edu", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("edu\_low", "edu\_medium", "edu\_high"),  
 ylab = "Inactivity", xlab = "Level of education")



oneway.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts0a2021, var.equal = FALSE)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: Inactivity\_rate and edu  
## F = 217, num df = 2.000, denom df = 49.094, p-value < 2.2e-16

pairwise.t.test(data\_anova\_nuts0a2021$Inactivity\_rate, data\_anova\_nuts0a2021$edu, p.adjust.method="bonferroni")

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: data\_anova\_nuts0a2021$Inactivity\_rate and data\_anova\_nuts0a2021$edu   
##   
## edu\_high edu\_low  
## edu\_low < 2e-16 -   
## edu\_medium 6.2e-12 < 2e-16  
##   
## P value adjustment method: bonferroni

We now reverse the analysis, i.e. we will analyze whether the year has an impact on the inactivity rate, all this for the 3 education levels.

We will remove the modality “edu total” like previously because there is no interest in having it.

mydata\_nuts0b <- mydata\_nuts0 %>%  
 filter (edu!="edu\_total")

We then divide our dataset in 3, for each level of education.

mydata\_nuts0blow <- mydata\_nuts0b %>%  
 filter (edu=="edu\_low")  
  
mydata\_nuts0bmedium <- mydata\_nuts0b %>%  
 filter (edu=="edu\_medium")  
  
mydata\_nuts0bhigh <- mydata\_nuts0b %>%  
 filter (edu=="edu\_high")

After this is completed, we can begin our analysis for each educational level.

# Low

To get started with our analysis, we can look for normality.

mydata\_nuts0blow\_2019 <- mydata\_nuts0blow %>%  
 filter (time=="2019")  
mydata\_nuts0blow\_2020 <- mydata\_nuts0blow %>%  
 filter (time=="2020")  
mydata\_nuts0blow\_2021 <- mydata\_nuts0blow %>%  
 filter (time=="2021")

shapiro.test(mydata\_nuts0blow\_2019$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0blow\_2019$Inactivity\_rate  
## W = 0.98474, p-value = 0.9053

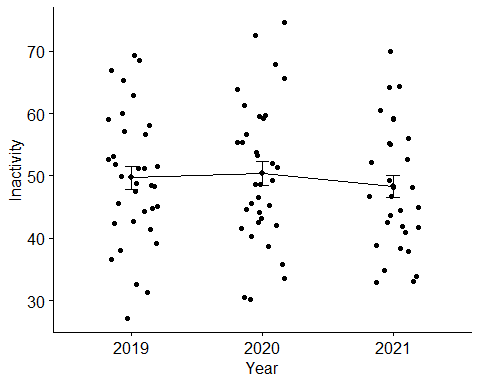
shapiro.test(mydata\_nuts0blow\_2020$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0blow\_2020$Inactivity\_rate  
## W = 0.98262, p-value = 0.8509

shapiro.test(mydata\_nuts0blow\_2021$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0blow\_2021$Inactivity\_rate  
## W = 0.96667, p-value = 0.4325

data\_anova\_nuts0blow <- mydata\_nuts0blow %>%  
 select(geo, time, Inactivity\_rate)  
  
ggline(data\_anova\_nuts0blow, x = "time", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("2019", "2020", "2021"),  
 ylab = "Inactivity", xlab = "Year")



oneway.test(Inactivity\_rate ~ time, data = data\_anova\_nuts0blow, var.equal = FALSE)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: Inactivity\_rate and time  
## F = 0.32804, num df = 2.000, denom df = 63.901, p-value = 0.7215

pairwise.t.test(data\_anova\_nuts0blow$Inactivity\_rate, data\_anova\_nuts0blow$time, p.adjust.method="bonferroni")

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: data\_anova\_nuts0blow$Inactivity\_rate and data\_anova\_nuts0blow$time   
##   
## 2019 2020  
## 2020 1 -   
## 2021 1 1   
##   
## P value adjustment method: bonferroni

To get started with our analysis, we can look for normality.

mydata\_nuts0bmedium\_2019 <- mydata\_nuts0bmedium %>%  
 filter (time=="2019")  
mydata\_nuts0bmedium\_2020 <- mydata\_nuts0bmedium %>%  
 filter (time=="2020")  
mydata\_nuts0bmedium\_2021 <- mydata\_nuts0bmedium %>%  
 filter (time=="2021")

shapiro.test(mydata\_nuts0bmedium\_2019$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0bmedium\_2019$Inactivity\_rate  
## W = 0.91742, p-value = 0.01367

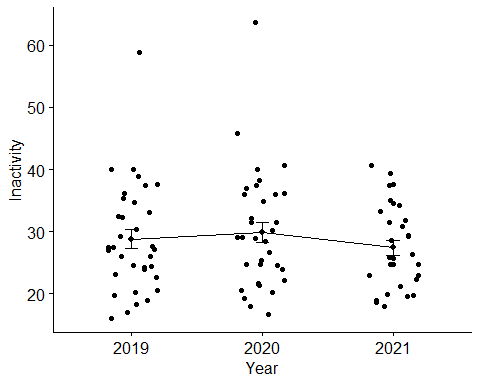
shapiro.test(mydata\_nuts0bmedium\_2020$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0bmedium\_2020$Inactivity\_rate  
## W = 0.90312, p-value = 0.005553

shapiro.test(mydata\_nuts0bmedium\_2021$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0bmedium\_2021$Inactivity\_rate  
## W = 0.94516, p-value = 0.1147

data\_anova\_nuts0bmedium <- mydata\_nuts0bmedium %>%  
 select(geo, time, Inactivity\_rate)  
  
ggline(data\_anova\_nuts0bmedium, x = "time", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("2019", "2020", "2021"),  
 ylab = "Inactivity", xlab = "Year")



oneway.test(Inactivity\_rate ~ time, data = data\_anova\_nuts0bmedium, var.equal = FALSE)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: Inactivity\_rate and time  
## F = 0.77318, num df = 2.000, denom df = 63.501, p-value = 0.4658

pairwise.t.test(data\_anova\_nuts0bmedium$Inactivity\_rate, data\_anova\_nuts0bmedium$time, p.adjust.method="bonferroni")

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: data\_anova\_nuts0bmedium$Inactivity\_rate and data\_anova\_nuts0bmedium$time   
##   
## 2019 2020  
## 2020 1.00 -   
## 2021 1.00 0.72  
##   
## P value adjustment method: bonferroni

To get started with our analysis, we can look for normality.

mydata\_nuts0bhigh\_2019 <- mydata\_nuts0blow %>%  
 filter (time=="2019")  
mydata\_nuts0bhigh\_2020 <- mydata\_nuts0bhigh %>%  
 filter (time=="2020")  
mydata\_nuts0bhigh\_2021 <- mydata\_nuts0bhigh %>%  
 filter (time=="2021")

shapiro.test(mydata\_nuts0bhigh\_2019$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0bhigh\_2019$Inactivity\_rate  
## W = 0.98474, p-value = 0.9053

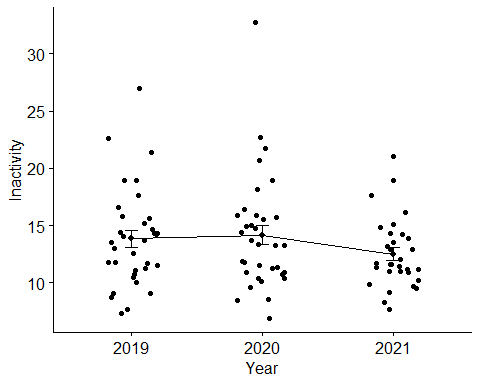
shapiro.test(mydata\_nuts0bhigh\_2020$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0bhigh\_2020$Inactivity\_rate  
## W = 0.87249, p-value = 0.0009294

shapiro.test(mydata\_nuts0bhigh\_2021$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts0bhigh\_2021$Inactivity\_rate  
## W = 0.9342, p-value = 0.05711

data\_anova\_nuts0bhigh <- mydata\_nuts0bhigh %>%  
 select(geo, time, Inactivity\_rate)  
  
ggline(data\_anova\_nuts0bhigh, x = "time", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("2019", "2020", "2021"),  
 ylab = "Inactivity", xlab = "Year")



out <- ag.test(Inactivity\_rate ~ time, data = data\_anova\_nuts0bhigh)

##   
## Alexander-Govern Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and time   
##   
## statistic : 3.459807   
## parameter : 2   
## p.value : 0.1773015   
##   
## Result : Difference is not statistically significant.   
## -------------------------------------------------------------

out2 <- bf.test(Inactivity\_rate ~ time, data = data\_anova\_nuts0bhigh)

##   
## Brown-Forsythe Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and time   
##   
## statistic : 1.395411   
## num df : 2   
## denom df : 86.97407   
## p.value : 0.2532215   
##   
## Result : Difference is not statistically significant.   
## -------------------------------------------------------------

one.way <- aov(Inactivity\_rate ~ time, data = data\_anova\_nuts0bhigh)  
summary(one.way)

## Df Sum Sq Mean Sq F value Pr(>F)  
## time 2 48.6 24.3 1.365 0.26  
## Residuals 96 1709.1 17.8

kruskal.test(Inactivity\_rate ~ time, data = data\_anova\_nuts0bhigh)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Inactivity\_rate by time  
## Kruskal-Wallis chi-squared = 2.2417, df = 2, p-value = 0.326

We begin by categorizing education levels. We only need to get rid of the mode “edu total” to accomplish this.

mydata\_nuts2a <- mydata\_nuts2 %>%  
 filter (edu!="edu\_total")

Then, for each year, we divide our dataset in 3.

mydata\_nuts2a2019 <- mydata\_nuts2a %>%  
 filter (time=="2019")  
  
mydata\_nuts2a2020 <- mydata\_nuts2a %>%  
 filter (time=="2020")  
  
mydata\_nuts2a2021 <- mydata\_nuts2a %>%  
 filter (time=="2021")

Once this is finished, we can begin each year’s analysis. To get started with our analysis, we can look for normality.

mydata\_nuts2a2019\_low <- mydata\_nuts2a2019 %>%  
 filter (edu=="edu\_low")  
mydata\_nuts2a2019\_medium <- mydata\_nuts2a2019 %>%  
 filter (edu=="edu\_medium")  
mydata\_nuts2a2019\_high <- mydata\_nuts2a2019 %>%  
 filter (edu=="edu\_high")

shapiro.test(mydata\_nuts2a2019\_low$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2a2019\_low$Inactivity\_rate  
## W = 0.9729, p-value = 4.302e-05

shapiro.test(mydata\_nuts2a2019\_medium$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2a2019\_medium$Inactivity\_rate  
## W = 0.87646, p-value = 3.659e-14

shapiro.test(mydata\_nuts2a2019\_high$Inactivity\_rate)

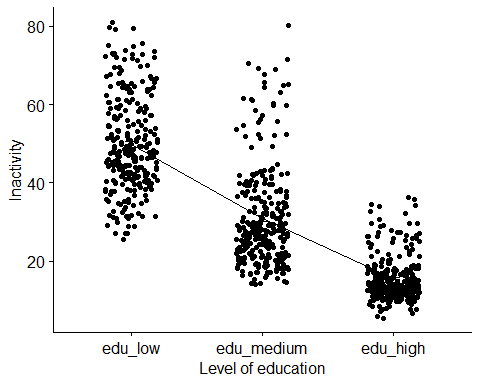
##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2a2019\_high$Inactivity\_rate  
## W = 0.87519, p-value = 3.052e-14

All pvalues are < 0,05, so we reject H0. We will use a non-parametric Anova because we cannot say that the sample follows a normal distribution. We can visualize our data.

data\_anova\_nuts2a2019 <- mydata\_nuts2a2019 %>%  
 select(geo, edu, Inactivity\_rate)  
  
ggline(data\_anova\_nuts2a2019, x = "edu", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("edu\_low", "edu\_medium", "edu\_high"),  
 ylab = "Inactivity", xlab = "Level of education")

## Warning: Removed 1 rows containing non-finite values (`stat\_summary()`).

## Warning: Removed 1 rows containing missing values (`geom\_point()`).



out <- ag.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2019)

##   
## Alexander-Govern Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and edu   
##   
## statistic : 855.3101   
## parameter : 2   
## p.value : 1.869729e-186   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

out2 <- bf.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2019)

##   
## Brown-Forsythe Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and edu   
##   
## statistic : 710.6454   
## num df : 2   
## denom df : 657.5822   
## p.value : 4.41155e-165   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

one.way <- aov(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2019)  
summary(one.way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## edu 2 162001 81000 711 <2e-16 \*\*\*  
## Residuals 827 94219 114   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 1 observation deleted due to missingness

kruskal.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2019)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Inactivity\_rate by edu  
## Kruskal-Wallis chi-squared = 583.14, df = 2, p-value < 2.2e-16

mydata\_nuts2a2020\_low <- mydata\_nuts2a2020 %>%  
 filter (edu=="edu\_low")  
mydata\_nuts2a2020\_medium <- mydata\_nuts2a2020 %>%  
 filter (edu=="edu\_medium")  
mydata\_nuts2a2020\_high <- mydata\_nuts2a2020 %>%  
 filter (edu=="edu\_high")

shapiro.test(mydata\_nuts2a2020\_low$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2a2020\_low$Inactivity\_rate  
## W = 0.98017, p-value = 0.001002

shapiro.test(mydata\_nuts2a2020\_medium$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2a2020\_medium$Inactivity\_rate  
## W = 0.86837, p-value = 1.171e-14

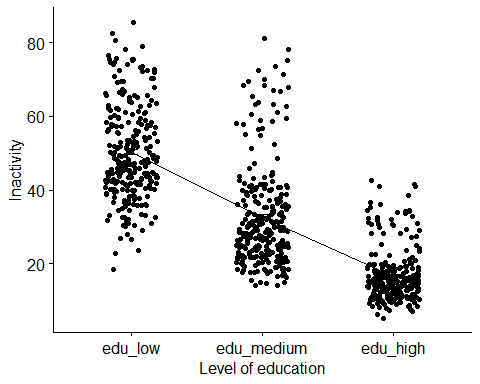
shapiro.test(mydata\_nuts2a2020\_high$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2a2020\_high$Inactivity\_rate  
## W = 0.83715, p-value < 2.2e-16

data\_anova\_nuts2a2020 <- mydata\_nuts2a2020 %>%  
 select(geo, edu, Inactivity\_rate)  
  
ggline(data\_anova\_nuts2a2020, x = "edu", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("edu\_low", "edu\_medium", "edu\_high"),  
 ylab = "Inactivity", xlab = "Level of education")

## Warning: Removed 14 rows containing non-finite values (`stat\_summary()`).

## Warning: Removed 14 rows containing missing values (`geom\_point()`).



out <- ag.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2020)

##   
## Alexander-Govern Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and edu   
##   
## statistic : 739.4059   
## parameter : 2   
## p.value : 2.754596e-161   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

out2 <- bf.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2020)

##   
## Brown-Forsythe Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and edu   
##   
## statistic : 574.6544   
## num df : 2   
## denom df : 674.06   
## p.value : 2.206654e-146   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

one.way <- aov(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2020)  
summary(one.way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## edu 2 153362 76681 577.7 <2e-16 \*\*\*  
## Residuals 814 108051 133   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 14 observations deleted due to missingness

kruskal.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2020)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Inactivity\_rate by edu  
## Kruskal-Wallis chi-squared = 542.86, df = 2, p-value < 2.2e-16

mydata\_nuts2a2021\_low <- mydata\_nuts2a2021 %>%  
 filter (edu=="edu\_low")  
mydata\_nuts2a2021\_medium <- mydata\_nuts2a2021 %>%  
 filter (edu=="edu\_medium")  
mydata\_nuts2a2021\_high <- mydata\_nuts2a2021 %>%  
 filter (edu=="edu\_high")

shapiro.test(mydata\_nuts2a2021\_low$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2a2021\_low$Inactivity\_rate  
## W = 0.96831, p-value = 2.869e-05

shapiro.test(mydata\_nuts2a2021\_medium$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2a2021\_medium$Inactivity\_rate  
## W = 0.93948, p-value = 9.544e-09

shapiro.test(mydata\_nuts2a2021\_high$Inactivity\_rate)

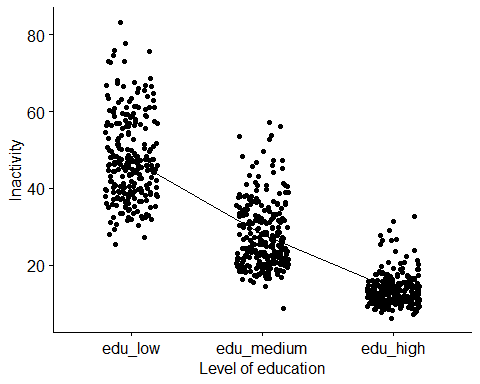
##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2a2021\_high$Inactivity\_rate  
## W = 0.88197, p-value = 3.604e-13

The 3 groups have a pvalue < 0.05, so we can reject h0. We will use a non-parametric Anova because we cannot say that the sample follows a normal distribution. We can visualize our data.

data\_anova\_nuts2a2021 <- mydata\_nuts2a2021 %>%  
 select(geo, edu, Inactivity\_rate)  
  
ggline(data\_anova\_nuts2a2021, x = "edu", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("edu\_low", "edu\_medium", "edu\_high"),  
 ylab = "Inactivity", xlab = "Level of education")

## Warning: Removed 10 rows containing non-finite values (`stat\_summary()`).

## Warning: Removed 10 rows containing missing values (`geom\_point()`).



out <- ag.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2021)

##   
## Alexander-Govern Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and edu   
##   
## statistic : 944.5853   
## parameter : 2   
## p.value : 7.689599e-206   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

out2 <- bf.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2021)

##   
## Brown-Forsythe Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and edu   
##   
## statistic : 1024.534   
## num df : 2   
## denom df : 536.918   
## p.value : 5.226298e-184   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

one.way <- aov(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2021)  
summary(one.way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## edu 2 145369 72684 1039 <2e-16 \*\*\*  
## Residuals 752 52587 70   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 10 observations deleted due to missingness

kruskal.test(Inactivity\_rate ~ edu, data = data\_anova\_nuts2a2020)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Inactivity\_rate by edu  
## Kruskal-Wallis chi-squared = 542.86, df = 2, p-value < 2.2e-16

We now return to the analysis of the years.

Since there is no interest in having the modality “edu total,” we will remove it as previously.

mydata\_nuts2b <- mydata\_nuts2 %>%  
 filter (edu!="edu\_total")

We then divide our dataset in 3, for each level of education.

mydata\_nuts2blow <- mydata\_nuts2b %>%  
 filter (edu=="edu\_low")  
  
mydata\_nuts2bmedium <- mydata\_nuts2b %>%  
 filter (edu=="edu\_medium")  
  
mydata\_nuts2bhigh <- mydata\_nuts2b %>%  
 filter (edu=="edu\_high")

After this is completed, we can begin our analysis for each educational level.

To get started with our analysis, we can look for normality.

mydata\_nuts2blow\_2019 <- mydata\_nuts2blow %>%  
 filter (time=="2019")  
mydata\_nuts2blow\_2020 <- mydata\_nuts2blow %>%  
 filter (time=="2020")  
mydata\_nuts2blow\_2021 <- mydata\_nuts2blow %>%  
 filter (time=="2021")

shapiro.test(mydata\_nuts2blow\_2019$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2blow\_2019$Inactivity\_rate  
## W = 0.9729, p-value = 4.302e-05

shapiro.test(mydata\_nuts2blow\_2020$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2blow\_2020$Inactivity\_rate  
## W = 0.98017, p-value = 0.001002

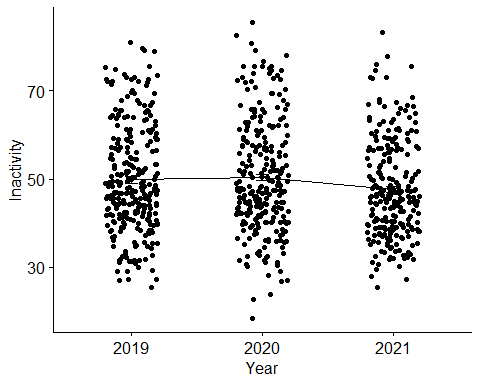
shapiro.test(mydata\_nuts2blow\_2021$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2blow\_2021$Inactivity\_rate  
## W = 0.96831, p-value = 2.869e-05

data\_anova\_nuts2blow <- mydata\_nuts2blow %>%  
 select(geo, time, Inactivity\_rate)  
  
ggline(data\_anova\_nuts2blow, x = "time", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("2019", "2020", "2021"),  
 ylab = "Inactivity", xlab = "Year")

## Warning: Removed 25 rows containing non-finite values (`stat\_summary()`).

## Warning: Removed 25 rows containing missing values (`geom\_point()`).



out <- ag.test(Inactivity\_rate ~ time, data = data\_anova\_nuts2blow)

##   
## Alexander-Govern Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and time   
##   
## statistic : 7.88114   
## parameter : 2   
## p.value : 0.01943713   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

out2 <- bf.test(Inactivity\_rate ~ time, data = data\_anova\_nuts2blow)

##   
## Brown-Forsythe Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and time   
##   
## statistic : 3.741774   
## num df : 2   
## denom df : 775.0994   
## p.value : 0.02414142   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

one.way <- aov(Inactivity\_rate ~ time, data = data\_anova\_nuts2blow)  
summary(one.way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## time 2 1059 529.3 3.722 0.0246 \*  
## Residuals 781 111042 142.2   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 25 observations deleted due to missingness

kruskal.test(Inactivity\_rate ~ time, data = data\_anova\_nuts2blow)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Inactivity\_rate by time  
## Kruskal-Wallis chi-squared = 7.3181, df = 2, p-value = 0.02576

To get started with our analysis, we can look for normality.

mydata\_nuts2bmedium\_2019 <- mydata\_nuts2bmedium %>%  
 filter (time=="2019")  
mydata\_nuts2bmedium\_2020 <- mydata\_nuts2bmedium %>%  
 filter (time=="2020")  
mydata\_nuts2bmedium\_2021 <- mydata\_nuts2bmedium %>%  
 filter (time=="2021")

shapiro.test(mydata\_nuts2bmedium\_2019$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2bmedium\_2019$Inactivity\_rate  
## W = 0.87646, p-value = 3.659e-14

shapiro.test(mydata\_nuts2bmedium\_2020$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2bmedium\_2020$Inactivity\_rate  
## W = 0.86837, p-value = 1.171e-14

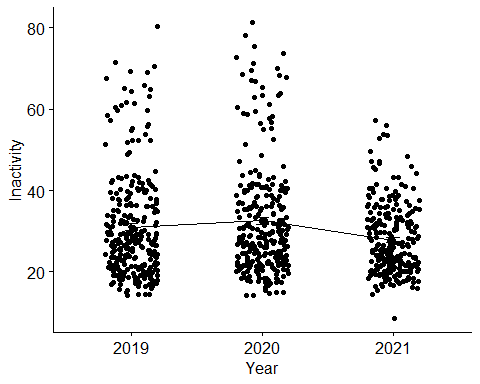
shapiro.test(mydata\_nuts2bmedium\_2021$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2bmedium\_2021$Inactivity\_rate  
## W = 0.93948, p-value = 9.544e-09

Here, all Pvalues are < 0,05 so we can reject h0.

We will use a non-parametric Anova because we cannot say that the sample follows a normal distribution. We can visualize our data.

data\_anova\_nuts2bmedium <- mydata\_nuts2bmedium %>%  
 select(geo, time, Inactivity\_rate)  
  
ggline(data\_anova\_nuts2bmedium, x = "time", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("2019", "2020", "2021"),  
 ylab = "Inactivity", xlab = "Year")



out <- ag.test(Inactivity\_rate ~ time, data = data\_anova\_nuts2bmedium)

##   
## Alexander-Govern Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and time   
##   
## statistic : 26.02098   
## parameter : 2   
## p.value : 2.236742e-06   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

out2 <- bf.test(Inactivity\_rate ~ time, data = data\_anova\_nuts2bmedium)

##   
## Brown-Forsythe Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and time   
##   
## statistic : 10.63924   
## num df : 2   
## denom df : 738.5805   
## p.value : 2.784486e-05   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

one.way <- aov(Inactivity\_rate ~ time, data = data\_anova\_nuts2bmedium)  
summary(one.way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## time 2 3018 1508.8 10.43 3.37e-05 \*\*\*  
## Residuals 806 116579 144.6   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

kruskal.test(Inactivity\_rate ~ time, data = data\_anova\_nuts2bmedium)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Inactivity\_rate by time  
## Kruskal-Wallis chi-squared = 12.3, df = 2, p-value = 0.002134

To get started with our analysis, we can look for normality.

mydata\_nuts2bhigh\_2019 <- mydata\_nuts2blow %>%  
 filter (time=="2019")  
mydata\_nuts2bhigh\_2020 <- mydata\_nuts2bhigh %>%  
 filter (time=="2020")  
mydata\_nuts2bhigh\_2021 <- mydata\_nuts2bhigh %>%  
 filter (time=="2021")

shapiro.test(mydata\_nuts2bhigh\_2019$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2bhigh\_2019$Inactivity\_rate  
## W = 0.9729, p-value = 4.302e-05

shapiro.test(mydata\_nuts2bhigh\_2020$Inactivity\_rate)

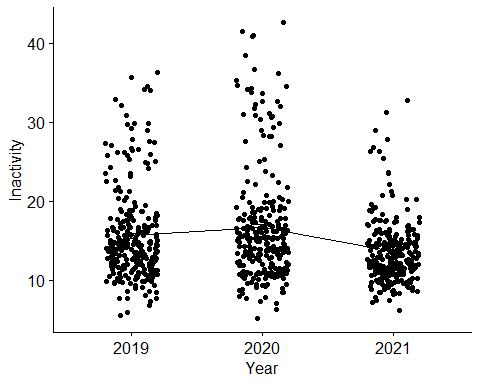
##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2bhigh\_2020$Inactivity\_rate  
## W = 0.83715, p-value < 2.2e-16

shapiro.test(mydata\_nuts2bhigh\_2021$Inactivity\_rate)

##   
## Shapiro-Wilk normality test  
##   
## data: mydata\_nuts2bhigh\_2021$Inactivity\_rate  
## W = 0.88197, p-value = 3.604e-13

Here, all Pvalues are < 0,05 so we can reject h0. We will use a non-parametric Anova because we cannot say that the sample follows a normal distribution. We can visualize our data.

data\_anova\_nuts2bhigh <- mydata\_nuts2bhigh %>%  
 select(geo, time, Inactivity\_rate)  
  
ggline(data\_anova\_nuts2bhigh, x = "time", y = "Inactivity\_rate",   
 add = c("mean\_se", "jitter"),   
 order = c("2019", "2020", "2021"),  
 ylab = "Inactivity", xlab = "Year")



out <- ag.test(Inactivity\_rate ~ time, data = data\_anova\_nuts2bhigh)

##   
## Alexander-Govern Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and time   
##   
## statistic : 44.28533   
## parameter : 2   
## p.value : 2.418596e-10   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

out2 <- bf.test(Inactivity\_rate ~ time, data = data\_anova\_nuts2bhigh)

##   
## Brown-Forsythe Test (alpha = 0.05)   
## -------------------------------------------------------------   
## data : Inactivity\_rate and time   
##   
## statistic : 18.79961   
## num df : 2   
## denom df : 707.7872   
## p.value : 1.10894e-08   
##   
## Result : Difference is statistically significant.   
## -------------------------------------------------------------

one.way <- aov(Inactivity\_rate ~ time, data = data\_anova\_nuts2bhigh)  
summary(one.way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## time 2 1245 622.3 18.42 1.51e-08 \*\*\*  
## Residuals 806 27235 33.8   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

kruskal.test(Inactivity\_rate ~ time, data = data\_anova\_nuts0bhigh)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Inactivity\_rate by time  
## Kruskal-Wallis chi-squared = 2.2417, df = 2, p-value = 0.326