Covid Effects on Institutional Performance

#Introduction:

How did Covid-19 affect institutional performance? A brief view of the education sector in Peru.

This project is a short and public version of an analysis I lead for the Ministry of Education of Peru (Minedu) in 2021, while my team and I designed the 2022 edition of a Results Based Financing (RBF) Mechanism called “Compromisos de Desempeño” (CdD or Performance Commitments).

The public education sector in Peru has the particularity of being divided into more than 220 decentralized institutions called “UGEL” (Unidades de Gestión Educativa Local) with similar functions and responsibilities for supplying public education throughout their individual jurisdictions. For further information on this topic, please refer to Article 73 of the Law 28044 (<http://www.minedu.gob.pe/p/ley_general_de_educacion_28044.pdf>). As part of the CdD Program, Minedu assesses various Key Performance Indicators for each UGEL, and the historical results from 2014 to the present day are publicly available on the CdD website (<http://www.minedu.gob.pe/cdd/>).

The next significant advantage is Peru was very open about their Covid data. The web “Datos Abiertos” (Open Data) allowed us to find every piece of data needed for the project. \* 2021 Peru population: [https://www.datosabiertos.gob.pe/dataset/población-peru](https://www.datosabiertos.gob.pe/dataset/poblaci%C3%B3n-peru) \* Covid related deaths: <https://www.datosabiertos.gob.pe/dataset/fallecidos-por-covid-19-ministerio-de-salud-minsa>

Bsed on the results, there was a significant impact of Covid-19 on UGELs performance in Peru in the first wave of 2020, particularly in areas that were hit harder by the pandemic. However, our analysis of the second wave in 2021 suggests that institutions may have adapted to the pandemic, and that there may have been underlying factors contributing to performance reduction.

## How to measure Institutional Performance?

Historical data of CdD results include the mean achievement value or “Valor Logrado” for each indicator in each UGEL or jurisdiction. These results can be analyzed by year and phase. For instance, during phase 1 of the CdD 2020, on January 31st, 2020, Minedu evaluated eight indicators among 223 institutions.

## How to measure the severity of the pandemic in each jurisdiction?

We decided to use the number of Covid-19 related deaths per 1,000 people in each jurisdiction for every quarter between 2020 and 2021. We believe that this data provides a more accurate representation of the impact of Covid-19 on the country, as Covid-19 infections may have been underreported due to the limited availability of testing, especially during the early stages of the pandemic.

# Preparation

Working Directory

f\_base <- "F:/Rodrigo/Antiguo Escritorio/Rodrigo/Cursos/Google Data Analytics/Course 8 Capstone/Project"  
  
setwd(f\_base)

##Data about schools in Peru

### Data from the Minedu (Ministerio de Educación)

Schools in Peru by district and UGEL jurisdiction - 2020 Cendo Educativo: <https://escale.minedu.gob.pe/uee/-/document_library_display/GMv7/view/6226837>

There are some filters I have to apply to clean this data \* ANEXO: It is higher than 0 if the row corresponds to a different building of the same school \* D\_FORMA: We need only schools and not other types of services, such as daycares \* NIV\_MOD: We need to filter out higher education from this analysis, as this institutions are under regional administration and are not inder the jurisdiction of the UGELs.

I generated a brief summary of the different levels of educations of Peru in the variable level\_of\_education.

#Levels of education  
peru\_schools\_full\_df %>%  
 group\_by(NIV\_MOD,D\_NIV\_MOD) %>%   
 summarise(n = n())

## `summarise()` has grouped output by 'NIV\_MOD'. You can override using the  
## `.groups` argument.

## # A tibble: 15 × 3  
## # Groups: NIV\_MOD [15]  
## NIV\_MOD D\_NIV\_MOD n  
## <chr> <chr> <int>  
## 1 A1 Inicial - Cuna 26  
## 2 A2 Inicial - JardÝn 33649  
## 3 A3 Inicial - Cuna-jardÝn 1661  
## 4 A5 Inicial - Programa no escolarizado 17942  
## 5 B0 Primaria 38411  
## 6 D1 Bßsica Alternativa-Inicial e Intermedio 852  
## 7 D2 Bßsica Alternativa-Avanzado 1532  
## 8 E0 Bßsica Especial 107  
## 9 E1 Bßsica Especial-Inicial 333  
## 10 E2 Bßsica Especial-Primaria 445  
## 11 F0 Secundaria 14942  
## 12 K0 Instituto Superior Pedag¾gico 178  
## 13 L0 TÚcnico Productiva 1653  
## 14 M0 Escuela Formaci¾n ArtÝstica 33  
## 15 T0 Instituto Superior Tecnol¾gico 785

#Data cleaning  
peru\_schools <- peru\_schools\_full\_df %>%   
 mutate(codmod = as.numeric(COD\_MOD),  
 codlocal = as.numeric(CODLOCAL),  
 ubigeo = as.numeric(CODGEO),  
 codooii = as.numeric(CODOOII),  
 level\_of\_education = case\_when(  
 NIV\_MOD %in% c("A1", "A2", "A3","A4") ~ "Preschool",  
 NIV\_MOD == "B0" ~ "Primary school",  
 NIV\_MOD == "F0" ~ "Secondary school",  
 NIV\_MOD %in% c("D1", "D2") ~ "Alternative school",  
 NIV\_MOD %in% c("E0","E1", "E2") ~ "Special needs school",  
 NIV\_MOD %in% c("K0","L0", "M0","T0") ~ "Higher education - not universities")) %>%   
 filter(ANEXO ==0,  
 D\_FORMA == "Escolarizada",  
 level\_of\_education != "Higher education - not universities") %>%   
 rename(school\_name = CEN\_EDU,  
 management = D\_GESTION,  
 urban\_context = DAREAMED,  
 geography = REGION\_NAT,  
 latitud = NLAT\_IE,  
 longitud = NLONG\_IE,  
 elevation\_msnm = ALTITUD) %>%   
 select(codmod,codlocal,ubigeo,codooii,level\_of\_education,school\_name,management,urban\_context,geography,latitud,longitud,elevation\_msnm)  
  
head(peru\_schools)

## codmod codlocal ubigeo codooii level\_of\_education  
## 1 415547 16100 20105 20001 Preschool  
## 2 415638 15172 20101 20001 Preschool  
## 3 415646 15186 20101 20001 Preschool  
## 4 415877 16751 20105 20001 Preschool  
## 5 567206 16119 20105 20001 Preschool  
## 6 567354 16124 20105 20001 Preschool  
## school\_name  
## 1 123  
## 2 122  
## 3 233  
## 4 COLEGIO PARROQUIAL NUESTRA SEÐORA DEL SAGRADO CORAZON DE JESUS  
## 5 268  
## 6 270  
## management urban\_context geography latitud longitud  
## 1 P·blica de gesti¾n directa Urbana SIERRA -9.51885 -77.53191  
## 2 P·blica de gesti¾n directa Urbana SIERRA -9.53067 -77.53196  
## 3 P·blica de gesti¾n directa Urbana SIERRA -9.53110 -77.52270  
## 4 Privada Urbana SIERRA -9.51667 -77.53148  
## 5 P·blica de gesti¾n directa Rural SIERRA -9.51394 -77.50403  
## 6 P·blica de gesti¾n directa Rural SIERRA -9.53546 -77.56759  
## elevation\_msnm  
## 1 3044  
## 2 3055  
## 3 3088  
## 4 3038  
## 5 3281  
## 6 3413

Now we need to relate each of Peru’s 1,874 districts to each of the 223 UGEL jurisdictions. To simplify this process, we will assign only one UGEL (with code “codooii”) to each district (with ubigeo code). In cases where a district falls within the jurisdiction of more than one UGEL, we will choose the UGEL with the highest number of schools in that district.

#Number of districts in Peru  
length(unique(peru\_schools$ubigeo))

## [1] 1874

#Number of UGEL in Peru  
length(unique(peru\_schools$codooii))

## [1] 223

#Ubigeo - Codooii relation  
ubigeo\_codooii <- peru\_schools %>%   
 group\_by(codooii,ubigeo) %>%   
 summarise(n\_schools = n())

## `summarise()` has grouped output by 'codooii'. You can override using the  
## `.groups` argument.

write\_xlsx(ubigeo\_codooii,"ubigeo\_codooii.xlsx")  
  
ubigeo\_codooii <- ubigeo\_codooii %>%  
 group\_by(ubigeo) %>%  
 arrange(desc(n\_schools)) %>%  
 slice(1) %>%  
 ungroup()

We will use the additional module of the Censo Educativo to collect the number of students enrolled in each school district.

students <- read.dbf("Matricula\_01.dbf", as.is = TRUE) %>%   
 as.data.frame() %>%   
 mutate(codmod = as.numeric(COD\_MOD),  
 codooii = as.numeric(CODOOII),  
 n\_students = rowSums(select(., starts\_with("D")))) %>%   
 group\_by(codmod, codooii) %>%   
 summarise(n\_students = sum(n\_students))

## `summarise()` has grouped output by 'codmod'. You can override using the  
## `.groups` argument.

head(students)

## # A tibble: 6 × 3  
## # Groups: codmod [6]  
## codmod codooii n\_students  
## <dbl> <dbl> <dbl>  
## 1 1506 150206 6  
## 2 1507 150206 9  
## 3 1509 150206 11  
## 4 1512 150206 7  
## 5 1513 150206 8  
## 6 1514 150206 6

Then we can calculate additional variables for each UGEL jurisdiction using the peru\_schools dataframe, such as: \* % rurality \* Mean elevation \* Number of schools \* Number of students \* sierra (Andes) \* selva (Rainforest)

I will use these variables in the linear regression models to evaluate whether the performance impact on the UGELs was more closely related to these characteristics rather than to COVID-19.

ugel\_vars <- merge(peru\_schools,students,by=c("codmod","codooii"), all.x = T) %>%  
 mutate(rural\_strudents = ifelse(urban\_context =="Rural",n\_students,0)) %>%   
 group\_by(codooii) %>%   
 summarise(n\_students = sum(n\_students),  
 rurality = sum(rural\_strudents)/sum(n\_students),  
 mean\_elevation\_msnm = mean(elevation\_msnm),  
 n\_schools = n(),  
 prcnt\_andes = sum(geography=="SIERRA")/n(),  
 prcnt\_selva = sum(geography=="SELVA")/n())  
head(ugel\_vars)

## # A tibble: 6 × 7  
## codooii n\_students rurality mean\_elevation\_msnm n\_schools prcnt\_andes  
## <dbl> <dbl> <dbl> <dbl> <int> <dbl>  
## 1 10001 15244 0.318 2354. 232 0.909   
## 2 10002 13257 0.360 762. 219 0   
## 3 10003 6921 0.543 1903. 122 0.328   
## 4 10004 27606 0.953 247. 479 0   
## 5 10005 12727 0.814 2153. 310 0.587   
## 6 10006 8820 0.764 1457. 193 0.0415  
## # ℹ 1 more variable: prcnt\_selva <dbl>

## Data about covid related deaths (ministerio de Salud)

<https://www.datosabiertos.gob.pe/dataset/fallecidos-por-covid-19-ministerio-de-salud-minsa> I’m going to group covid deaths by quarter and year for simplicity in the analysis.

covid\_deaths <- read.csv("fallecidos\_covid.csv",sep = ";") %>%   
 as.data.frame() %>%   
 mutate(year = as.integer(substr(FECHA\_FALLECIMIENTO, 1, 4)),  
 month = as.integer(substr(FECHA\_FALLECIMIENTO, 5, 6)),  
 ubigeo = as.numeric(UBIGEO),  
 age = EDAD\_DECLARADA) %>%   
 filter(!is.na(ubigeo))

## Warning: There was 1 warning in `mutate()`.  
## ℹ In argument: `ubigeo = as.numeric(UBIGEO)`.  
## Caused by warning:  
## ! NAs introducidos por coerción

covid\_deaths\_ubigeo <- covid\_deaths %>%   
 group\_by(ubigeo) %>%   
 summarise(covid\_deaths\_2020\_1 = sum(year==2020 & month <=3),  
 covid\_deaths\_2020\_2 = sum(year==2020 & month >3 & month<=6),  
 covid\_deaths\_2020\_3 = sum(year==2020 & month >6 & month<=9),  
 covid\_deaths\_2020\_4 = sum(year==2020 & month >9),  
 covid\_deaths\_2021\_1 = sum(year==2021 & month <=3),  
 covid\_deaths\_2021\_2 = sum(year==2021 & month >3 & month<=6),  
 covid\_deaths\_2021\_3 = sum(year==2021 & month >6 & month<=9),  
 covid\_deaths\_2021\_4 = sum(year==2021 & month >9),  
 covid\_deaths = n())  
head(covid\_deaths)

## FECHA\_CORTE FECHA\_FALLECIMIENTO EDAD\_DECLARADA SEXO CLASIFICACION\_DEF  
## 1 20230413 20210512 88 FEMENINO Criterio SINADEF  
## 2 20230413 20210624 58 MASCULINO Criterio SINADEF  
## 3 20230413 20210426 84 MASCULINO Criterio serolÃ³gico  
## 4 20230413 20210315 64 MASCULINO Criterio virolÃ³gico  
## 5 20230413 20210504 62 MASCULINO Criterio SINADEF  
## 6 20230413 20210321 61 FEMENINO Criterio SINADEF  
## DEPARTAMENTO PROVINCIA DISTRITO UBIGEO UUID year month ubigeo  
## 1 AREQUIPA AREQUIPA PAUCARPATA 040112 13889730 2021 5 40112  
## 2 AREQUIPA AREQUIPA PAUCARPATA 040112 13889749 2021 6 40112  
## 3 LA LIBERTAD SANCHEZ CARRION CHUGAY 130902 13889932 2021 4 130902  
## 4 LA LIBERTAD SANTIAGO DE CHUCO CACHICADAN 131003 13889956 2021 3 131003  
## 5 LIMA HUARAL CHANCAY 150605 13890034 2021 5 150605  
## 6 LIMA HUARAL HUARAL 150601 13890061 2021 3 150601  
## age  
## 1 88  
## 2 58  
## 3 84  
## 4 64  
## 5 62  
## 6 61

total\_deaths <- as.numeric(sum(covid\_deaths$covid\_deaths))  
print(total\_deaths)

## [1] 0

timeline\_covid <- covid\_deaths %>%   
 group\_by(year,month) %>%   
 summarise(covid\_deaths = n())

## `summarise()` has grouped output by 'year'. You can override using the  
## `.groups` argument.

## Distric population in 2021 (Estimation by the INEI)

<https://www.datosabiertos.gob.pe/dataset/población-peru>

Minedu uses ubigeo\_inei Minsa uses ubigeo\_inei

peru\_population <- read.csv("TB\_POBLACION\_INEI.csv",sep = ",") %>%   
 as.data.frame() %>%   
 mutate(ubigeo = ubigeo\_inei) %>%   
 group\_by(ubigeo) %>%   
 summarise(population = sum(Cantidad))  
head(peru\_population)

## # A tibble: 6 × 2  
## ubigeo population  
## <int> <int>  
## 1 10101 29041  
## 2 10102 291  
## 3 10103 1639  
## 4 10104 595  
## 5 10105 716  
## 6 10106 2088

sum(peru\_population$population)

## [1] 32526084

## Compromisos de Desempeño (CdD) results

CdD are evaluated in phases, each phase has an evaluation date. Also we need to keep only the results of UGEL using the type of entity column, but keeping the Regional Office of Callao (DRE CALLAO), because it functions as another UGEL.

#The source of the phases dates comes from the legal documents and directives on the web page  
phases\_cdd <- read\_excel("cdd\_phase\_date.xlsx") %>%   
 as.data.frame()  
  
#Results as available on the web page (validation made in-house by the team)  
cdd\_results <- read.csv("resultados\_hist\_cdd.csv") %>%   
 as.data.frame() %>%   
 filter(tipo\_entidad == "UGEL EJECUTORA" | tipo\_entidad =="UGEL OPERATIVA" | iged =="DRE CALLAO")  
cdd\_results <- merge(cdd\_results,phases\_cdd,by = c("periodo","tramo"))  
head(cdd\_results)

## periodo tramo indicador codooii tipo\_entidad iged  
## 1 2018 1 1.1 100010 UGEL EJECUTORA UGEL PACHITEA  
## 2 2018 1 1.1 120004 UGEL EJECUTORA UGEL CHANCHAMAYO  
## 3 2018 1 1.1 130012 UGEL EJECUTORA UGEL GRAN CHIMU  
## 4 2018 1 1.1 90006 UGEL EJECUTORA UGEL HUAYTARA  
## 5 2018 1 1.1 100004 UGEL EJECUTORA UGEL LAURICOCHA  
## 6 2018 1 1.1 80002 UGEL EJECUTORA UGEL ACOMAYO  
## ue region meta  
## 1 304-1386: GOB. REG. HUANUCO - EDUCACION PACHITEA HUANUCO 0.85  
## 2 303-1459: GOB. REG. DE JUNIN- EDUCACION CHANCHAMAYO JUNIN 0.85  
## 3 304-838: REGION LA LIBERTAD-EDUCACION GRAN CHIMU LA LIBERTAD 0.85  
## 4 311-1640: GOB. REG. HUANCAVELICA - UGEL HUAYTARA HUANCAVELICA 0.85  
## 5 309-1541: GOB. REG. HUANUCO - EDUCACION UGEL LAURICOCHA HUANUCO 0.80  
## 6 314-1668: GOB. REG. DPTO. CUSCO - EDUCACION ACOMAYO CUSCO 0.80  
## numerador denominador valor\_logrado\_ufd evaluacion cumplimiento\_ufd  
## 1 NA NA 0.96 1 1  
## 2 NA NA 0.95 1 1  
## 3 NA NA 0.93 1 1  
## 4 NA NA 0.85 1 1  
## 5 NA NA 0.83 1 1  
## 6 NA NA 1.00 1 1  
## compromiso cod\_ind\_interanual fecha\_corte mes\_corte  
## 1 1 D-01 2018-05-31 5  
## 2 1 D-01 2018-05-31 5  
## 3 1 D-01 2018-05-31 5  
## 4 1 D-01 2018-05-31 5  
## 5 1 D-01 2018-05-31 5  
## 6 1 D-01 2018-05-31 5

## Timeline of covid and CdD

timeline\_cdd <- cdd\_results %>%   
 filter(periodo > 2019) %>%   
 mutate(year = periodo,  
 month = mes\_corte) %>%   
 group\_by(year,month) %>%   
 summarise(mean\_vl = mean(valor\_logrado\_ufd,na.rm = T))

## `summarise()` has grouped output by 'year'. You can override using the  
## `.groups` argument.

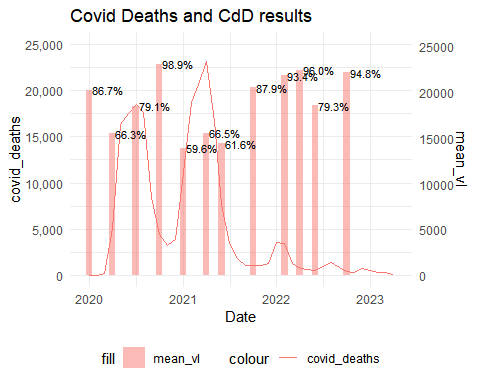
head(timeline\_cdd)

## # A tibble: 6 × 3  
## # Groups: year [2]  
## year month mean\_vl  
## <int> <dbl> <dbl>  
## 1 2020 1 0.867  
## 2 2020 4 0.663  
## 3 2020 7 0.791  
## 4 2020 10 0.989  
## 5 2021 1 0.596  
## 6 2021 4 0.665

#Timeline  
timeline <- merge(timeline\_covid, timeline\_cdd, by = c("year","month"),all = T) %>%   
 add\_row(year = 2020, month = 2) %>%   
 mutate(date = ymd(paste(year, month, "01", sep = "-")),  
 mean\_vl = ifelse(is.na(mean\_vl),0,mean\_vl),  
 covid\_deaths = ifelse(is.na(covid\_deaths),0,covid\_deaths))  
head(timeline)

## year month covid\_deaths mean\_vl date  
## 1 2020 1 0 0.8670502 2020-01-01  
## 2 2020 3 181 0.0000000 2020-03-01  
## 3 2020 4 5131 0.6634098 2020-04-01  
## 4 2020 5 16364 0.0000000 2020-05-01  
## 5 2020 6 17462 0.0000000 2020-06-01  
## 6 2020 7 18501 0.7910145 2020-07-01

timelinechart <- ggplot(timeline, aes(x = date)) +  
 geom\_line(aes(y = covid\_deaths, color = "covid\_deaths")) +  
 geom\_bar(aes(y = mean\_vl \* max(covid\_deaths), fill = "mean\_vl"), stat = "identity", alpha = 0.5) +  
 scale\_y\_continuous(name = "covid\_deaths", limits = c(0, 25000), sec.axis = sec\_axis(~./max(timeline$mean\_vl), name = "mean\_vl"), labels = scales::comma) +  
 labs(x = "Date", title = "Covid Deaths and CdD results") +  
 theme\_minimal() +  
 theme(legend.position = "bottom") +  
 geom\_text(data = subset(timeline, mean\_vl > 0),  
 aes(y = mean\_vl \* max(covid\_deaths), label = paste0(scales::percent(mean\_vl, accuracy = 0.1))),  
 hjust = -0.1, size = 3, color = "black")  
timelinechart



Hits in UGELs performance on the CdD indicators coincide with the Covid-19 waves in Peru. The first Covid-19 case in Peru was reported in March 2020, by this time, Minedu had already evaluated 8 indicators in the phase 1 of the CdD 2020. This event, gives us the possibility to see how the UGELs would have performed on normal conditions. The Pandemic hit Peru particularly hard in the 2nd and 3rd quarters of 2020, and then again in the 1st and 2nd quarters of 2021.

We want to analyze how these tough quarters affected the performance of UGELs, measured as achievement value. \* Correlation of the mean achievement value (phases 2 and 3 of 2020) vs Covid-19 related deaths (quarters 2 and 3 of 2020). \* Correlation of the mean achievement value (phases 1 and 2 of 2021) Covid-19 related deaths (quarters 1 and 2 of 2021). \* Covid hit: Correlation and linear regression of (phase 1 2020 - mean vachievement value of phases 2 and 3) vs Covid-19 related deaths (quarters 2 and 3 of 2020).

## Covid deaths in each UGEL jurisdiction, by 1,000 inhabitants

* covid\_deaths
* peru\_population
* ubigeo\_codooii

#First We have to merge all three dataframes by ubigeo  
df\_list <- list(covid\_deaths\_ubigeo, peru\_population, ubigeo\_codooii)  
covid\_deaths\_pops <- df\_list %>% reduce(full\_join, by="ubigeo")  
  
#Now we need to calculate how many Covid-19 related death occurred in every UGEL jurisdiction, as well as the population of each jurisdiction  
# Also we want to analyze covid deaths relative to jurisdiction population. I chose covid deaths per 1k inhabitants.  
covid\_deaths\_ugel <- covid\_deaths\_pops %>%   
 group\_by(codooii) %>%   
 summarise(covid\_deaths\_2020\_1 = 1000\*sum(covid\_deaths\_2020\_1,na.rm = T)/sum(population),  
 covid\_deaths\_2020\_2 = 1000\*sum(covid\_deaths\_2020\_2,na.rm = T)/sum(population),  
 covid\_deaths\_2020\_3 = 1000\*sum(covid\_deaths\_2020\_3,na.rm = T)/sum(population),  
 covid\_deaths\_2020\_4 = 1000\*sum(covid\_deaths\_2020\_4,na.rm = T)/sum(population),  
 covid\_deaths\_2021\_1 = 1000\*sum(covid\_deaths\_2021\_1,na.rm = T)/sum(population),  
 covid\_deaths\_2021\_2 = 1000\*sum(covid\_deaths\_2021\_2,na.rm = T)/sum(population),  
 covid\_deaths\_2021\_3 = 1000\*sum(covid\_deaths\_2021\_3,na.rm = T)/sum(population),  
 covid\_deaths\_2021\_4 = 1000\*sum(covid\_deaths\_2021\_4,na.rm = T)/sum(population),  
 covid\_deaths = sum(covid\_deaths,na.rm = T),  
 population = sum(population))  
  
head(covid\_deaths\_ugel)

## # A tibble: 6 × 11  
## codooii covid\_deaths\_2020\_1 covid\_deaths\_2020\_2 covid\_deaths\_2020\_3  
## <dbl> <dbl> <dbl> <dbl>  
## 1 10001 0 0.143 0.681  
## 2 10002 0 0.600 3.15   
## 3 10003 0 0.0584 0.409  
## 4 10004 0 0.360 0.936  
## 5 10005 0 0.0955 0.191  
## 6 10006 0 0 0.159  
## # ℹ 7 more variables: covid\_deaths\_2020\_4 <dbl>, covid\_deaths\_2021\_1 <dbl>,  
## # covid\_deaths\_2021\_2 <dbl>, covid\_deaths\_2021\_3 <dbl>,  
## # covid\_deaths\_2021\_4 <dbl>, covid\_deaths <int>, population <int>

CdD data collapsed at the UGEL level (change over year) Merge ugel\_vars + covid\_deaths\_ugel + cdd\_results Data needed: \* Mean achievement value by year and phase from 2020 to 2021 \* Identify common indicators between 2019 - 2021

# Mean achievement value by year and phase from 2020 to 2021  
cdd\_results\_ugel <- cdd\_results %>%   
 filter(periodo == 2020 | periodo == 2021) %>%   
 group\_by(codooii) %>%   
 summarise(ugel = first(iged),  
 region = first(region),  
 vl\_2020\_1 = mean(valor\_logrado\_ufd[periodo==2020 & tramo==1],na.rm = T),  
 vl\_2020\_2 = mean(valor\_logrado\_ufd[periodo==2020 & tramo==2],na.rm = T),  
 vl\_2020\_2\_3 = mean(valor\_logrado\_ufd[periodo==2020 & (tramo==2 | tramo==3)],na.rm = T),  
 vl\_2020\_4 = mean(valor\_logrado\_ufd[periodo==2020 & tramo==4],na.rm = T),  
 vl\_2021\_1\_2 = mean(valor\_logrado\_ufd[periodo==2021 & (tramo==1 | tramo==2)],na.rm = T),  
 vl\_2021\_3 = mean(valor\_logrado\_ufd[periodo==2021 & tramo==3],na.rm = T))  
head(cdd\_results\_ugel)

## # A tibble: 6 × 9  
## codooii ugel region vl\_2020\_1 vl\_2020\_2 vl\_2020\_2\_3 vl\_2020\_4 vl\_2021\_1\_2  
## <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 10001 UGEL CHA… AMAZO… 0.838 0.37 0.620 NaN 0.725  
## 2 10002 UGEL BAG… AMAZO… 0.923 0.675 0.721 0.969 0.860  
## 3 10003 UGEL BON… AMAZO… 0.919 0.368 0.626 NaN 0.703  
## 4 10004 UGEL CON… AMAZO… 0.767 0.701 0.689 0.986 0.546  
## 5 10005 UGEL LUYA AMAZO… 0.868 0.388 0.639 NaN 0.4   
## 6 10006 UGEL ROD… AMAZO… 0.875 0.368 0.627 NaN 0.746  
## # ℹ 1 more variable: vl\_2021\_3 <dbl>

#Analysis

#Dataframe  
head(cdd\_results\_ugel)

## # A tibble: 6 × 9  
## codooii ugel region vl\_2020\_1 vl\_2020\_2 vl\_2020\_2\_3 vl\_2020\_4 vl\_2021\_1\_2  
## <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 10001 UGEL CHA… AMAZO… 0.838 0.37 0.620 NaN 0.725  
## 2 10002 UGEL BAG… AMAZO… 0.923 0.675 0.721 0.969 0.860  
## 3 10003 UGEL BON… AMAZO… 0.919 0.368 0.626 NaN 0.703  
## 4 10004 UGEL CON… AMAZO… 0.767 0.701 0.689 0.986 0.546  
## 5 10005 UGEL LUYA AMAZO… 0.868 0.388 0.639 NaN 0.4   
## 6 10006 UGEL ROD… AMAZO… 0.875 0.368 0.627 NaN 0.746  
## # ℹ 1 more variable: vl\_2021\_3 <dbl>

head(covid\_deaths\_ugel)

## # A tibble: 6 × 11  
## codooii covid\_deaths\_2020\_1 covid\_deaths\_2020\_2 covid\_deaths\_2020\_3  
## <dbl> <dbl> <dbl> <dbl>  
## 1 10001 0 0.143 0.681  
## 2 10002 0 0.600 3.15   
## 3 10003 0 0.0584 0.409  
## 4 10004 0 0.360 0.936  
## 5 10005 0 0.0955 0.191  
## 6 10006 0 0 0.159  
## # ℹ 7 more variables: covid\_deaths\_2020\_4 <dbl>, covid\_deaths\_2021\_1 <dbl>,  
## # covid\_deaths\_2021\_2 <dbl>, covid\_deaths\_2021\_3 <dbl>,  
## # covid\_deaths\_2021\_4 <dbl>, covid\_deaths <int>, population <int>

head(ugel\_vars)

## # A tibble: 6 × 7  
## codooii n\_students rurality mean\_elevation\_msnm n\_schools prcnt\_andes  
## <dbl> <dbl> <dbl> <dbl> <int> <dbl>  
## 1 10001 15244 0.318 2354. 232 0.909   
## 2 10002 13257 0.360 762. 219 0   
## 3 10003 6921 0.543 1903. 122 0.328   
## 4 10004 27606 0.953 247. 479 0   
## 5 10005 12727 0.814 2153. 310 0.587   
## 6 10006 8820 0.764 1457. 193 0.0415  
## # ℹ 1 more variable: prcnt\_selva <dbl>

#We need to merge these three dataframes by codooii  
df\_list2 <- list(cdd\_results\_ugel, covid\_deaths\_ugel, ugel\_vars)  
analysis\_df <- df\_list2 %>% reduce(full\_join, by="codooii")  
head(analysis\_df)

## # A tibble: 6 × 25  
## codooii ugel region vl\_2020\_1 vl\_2020\_2 vl\_2020\_2\_3 vl\_2020\_4 vl\_2021\_1\_2  
## <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 10001 UGEL CHA… AMAZO… 0.838 0.37 0.620 NaN 0.725  
## 2 10002 UGEL BAG… AMAZO… 0.923 0.675 0.721 0.969 0.860  
## 3 10003 UGEL BON… AMAZO… 0.919 0.368 0.626 NaN 0.703  
## 4 10004 UGEL CON… AMAZO… 0.767 0.701 0.689 0.986 0.546  
## 5 10005 UGEL LUYA AMAZO… 0.868 0.388 0.639 NaN 0.4   
## 6 10006 UGEL ROD… AMAZO… 0.875 0.368 0.627 NaN 0.746  
## # ℹ 17 more variables: vl\_2021\_3 <dbl>, covid\_deaths\_2020\_1 <dbl>,  
## # covid\_deaths\_2020\_2 <dbl>, covid\_deaths\_2020\_3 <dbl>,  
## # covid\_deaths\_2020\_4 <dbl>, covid\_deaths\_2021\_1 <dbl>,  
## # covid\_deaths\_2021\_2 <dbl>, covid\_deaths\_2021\_3 <dbl>,  
## # covid\_deaths\_2021\_4 <dbl>, covid\_deaths <int>, population <int>,  
## # n\_students <dbl>, rurality <dbl>, mean\_elevation\_msnm <dbl>,  
## # n\_schools <int>, prcnt\_andes <dbl>, prcnt\_selva <dbl>

###Correlation: mean achievement value (phases 2 and 3 of CdD 2020) vs Covid-19 related deaths (quarters 2 and 3 of 2020). Minedu evaluated 17 indicators in phases 2 and 3 of CdD 2020.

Based on the results, it is apparent that during the first wave of Covid-19 in Peru, there was a strong correlation between the number of Covid-19 related deaths relative to the population of each jurisdiction, and the performance of the UGELs on the CdD Program of Minedu. This means that in jurisdictions that were hit harder by the pandemic, the performance of the UGELs was significantly lower. This could be due to a variety of factors, such as increased pressure on UGELs to respond to the pandemic, including adapting to remote learning models and ensuring the safety of students and staff, as well as dealing with the social and economic consequences of the pandemic.

It’s important to note that this correlation may not necessarily imply causation. Other factors, such as limited human capital in rural areas of the country and geographic challenges, may also have contributed to the variation in UGEL performance. Nonetheless, the correlation does suggest that the pandemic had a significant impact on the performance of UGELs in Peru.

#2020 subset  
df\_2020 <- analysis\_df %>%   
 mutate(covid\_deaths\_2020\_2\_3 = covid\_deaths\_2020\_2 + covid\_deaths\_2020\_3) %>%   
 filter(!is.na(vl\_2020\_2\_3)) %>%   
 select(vl\_2020\_2\_3,covid\_deaths\_2020\_2\_3)  
  
#Correlation test  
cor(df\_2020)

## vl\_2020\_2\_3 covid\_deaths\_2020\_2\_3  
## vl\_2020\_2\_3 1.000000 -0.246425  
## covid\_deaths\_2020\_2\_3 -0.246425 1.000000

cor\_2020 <- cor.test(df\_2020$covid\_deaths\_2020\_2\_3, df\_2020$vl\_2020\_2\_3)  
cor\_2020$estimate # correlation coefficient

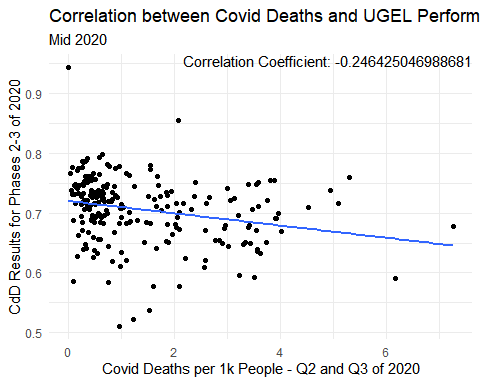
## cor   
## -0.246425

cor\_2020$p.value # p-value

## [1] 0.0002157563

#Visualization  
corr\_2020\_chart <- df\_2020 %>%   
 ggplot(aes(x = covid\_deaths\_2020\_2\_3,y=vl\_2020\_2\_3)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 labs(x = "Covid Deaths per 1k People - Q2 and Q3 of 2020",   
 y = "CdD Results for Phases 2-3 of 2020",   
 title = "Correlation between Covid Deaths and UGEL Performance", subtitle = "Mid 2020") +  
 annotate("text", x = Inf, y = Inf, label = paste0("Correlation Coefficient: ", cor\_2020$estimate), hjust = 1, vjust = 1) +  
 theme\_minimal()  
corr\_2020\_chart

## `geom\_smooth()` using formula = 'y ~ x'



###Correlation: mean achievement value (phases 1 and 2 of 2021) covid deaths (quarters 1 and 2 of 2021). Minedu evaluated 13 indicators in phases 1 and 2 of CdD 2021.

Based on the information provided, it seems that the second wave of Covid-19 related deaths in Peru may not be significantly correlated with UGEL performance, as the p-value is higher than .25. This could be an indication that institutions have become better adapted to the pandemic after a year of experience, and other factors may be contributing to the performance hit.

#2021 subset  
df\_2021 <- analysis\_df %>%   
 mutate(covid\_deaths\_2021\_1\_2 = covid\_deaths\_2021\_1 + covid\_deaths\_2021\_2) %>%   
 filter(!is.na(vl\_2021\_1\_2)) %>%   
 select(vl\_2021\_1\_2,covid\_deaths\_2021\_1\_2)  
  
#Correlation test  
cor(df\_2021)

## vl\_2021\_1\_2 covid\_deaths\_2021\_1\_2  
## vl\_2021\_1\_2 1.00000000 0.07438889  
## covid\_deaths\_2021\_1\_2 0.07438889 1.00000000

cor\_2021 <- cor.test(df\_2021$covid\_deaths\_2021\_1\_2, df\_2021$vl\_2021\_1\_2)  
cor\_2021$estimate # correlation coefficient

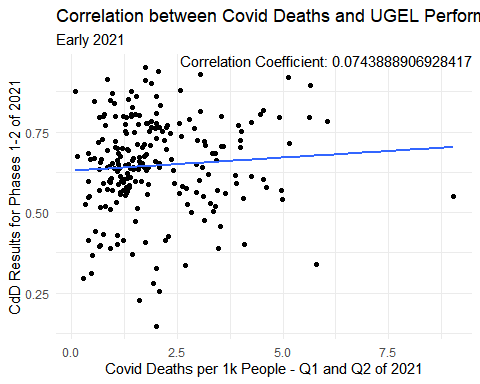
## cor   
## 0.07438889

cor\_2021$p.value # p-value

## [1] 0.2686601

#Visualization  
corr\_2021\_chart <- df\_2021 %>%   
 ggplot(aes(x = covid\_deaths\_2021\_1\_2,y=vl\_2021\_1\_2)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 labs(x = "Covid Deaths per 1k People - Q1 and Q2 of 2021",   
 y = "CdD Results for Phases 1-2 of 2021",   
 title = "Correlation between Covid Deaths and UGEL Performance", subtitle = "Early 2021") +  
 annotate("text", x = Inf, y = Inf, label = paste0("Correlation Coefficient: ", cor\_2021$estimate), hjust = 1, vjust = 1) +  
 theme\_minimal()  
corr\_2021\_chart

## `geom\_smooth()` using formula = 'y ~ x'



###Covid hit: Correlation of Performance hit (mean achievement value during phases 2 and 3 of 2020 minus mean achievement value in phase 1) vs covid deaths (quarters 2 and 3 of 2020).

Based on the analysis conducted, it appears that there was a correlation between the impact of Covid-19 on each jurisdiction and the reduced performance of the UGELs in those jurisdictions during the first wave of the pandemic.

#2021 subset  
df\_2020\_hit <- analysis\_df %>%   
 mutate(covid\_deaths\_2020\_2\_3 = covid\_deaths\_2020\_2 + covid\_deaths\_2020\_3,  
 vl\_2020\_dif = vl\_2020\_2\_3 - vl\_2020\_1) %>%   
 filter(!is.na(vl\_2020\_dif)) %>%   
 select(vl\_2020\_dif,covid\_deaths\_2020\_2\_3)  
  
#Correlation test  
cor(df\_2020\_hit)

## vl\_2020\_dif covid\_deaths\_2020\_2\_3  
## vl\_2020\_dif 1.0000000 -0.1068479  
## covid\_deaths\_2020\_2\_3 -0.1068479 1.0000000

cor\_2020\_hit <- cor.test(df\_2020\_hit$covid\_deaths\_2020\_2\_3, df\_2020\_hit$vl\_2020\_dif)  
cor\_2020\_hit$estimate # correlation coefficient

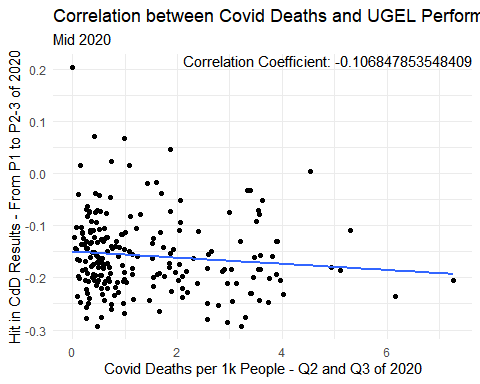
## cor   
## -0.1068479

cor\_2020\_hit$p.value # p-value

## [1] 0.113208

#Visualization  
corr\_2020hit\_chart <- df\_2020\_hit %>%   
 ggplot(aes(x = covid\_deaths\_2020\_2\_3,y=vl\_2020\_dif)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 labs(x = "Covid Deaths per 1k People - Q2 and Q3 of 2020",   
 y = "Hit in CdD Results - From P1 to P2-3 of 2020",   
 title = "Correlation between Covid Deaths and UGEL Performance Change", subtitle = "Mid 2020") +  
 annotate("text", x = Inf, y = Inf, label = paste0("Correlation Coefficient: ", cor\_2020\_hit$estimate), hjust = 1, vjust = 1) +  
 theme\_minimal()  
corr\_2020hit\_chart

## `geom\_smooth()` using formula = 'y ~ x'



### Linear regressions

Linear regressions are a powerful statistical tool that can help analyze the impact of other factors, other than Covid-19, that affected UGELs performance on the CdD program. By using regression models, we can estimate the relationships between different variables in different jurisdictions. This approach can provide a more comprehensive analysis of the underlying factors that influenced institutional performance during the pandemic.

In order to estimate the performance of UGELs using linear regression models, we included several variables in the analysis. Besides Covid-19 related deaths, which were measured relative to the population in each quarter, we also considered the number of students in each jurisdiction, the degree of rurality, the number of schools, the mean elevation, and the percentage of schools located in the Andes or in the rain forest. By analyzing the impact of these different factors, we can better understand how they have influenced the performance of UGELs during the pandemic. However, only few factors had an effect on UGELs performance, here’s a brief overview of the analysis.

In the first model, we can observe that Covid-19 related deaths had a significant impact on UGELs performance during the first wave in 2020, but the effect was reduced during the third quarter of that year. Additionally, the number of students had a significant impact on CdD results, probably due to the institutional stress caused by a higher population.

In the second model, we found that the impact of Covid-19 on UGELs performance during the second wave in Peru was not as significant as during the first wave in 2020. However, the model revealed some interesting and somewhat contradictory results, as it showed that Covid-19 related deaths in the 2nd quarter and rurality had a positive impact on UGELs performance. One possible explanation could be that urban areas had multiple lockdowns mandates enforced by the government relative to rural areas. However it is important to conduct more research to fully understand the underlying factors contributing to these results.

Finally, the third model may indicate that, although this analysis may be headed on the right path, further analysis may be needed to finally understand how Covid-19 effected UGELs performance. None of the variables selected explain the reduction in achievement value between phase 1 of the CdD 2020 and phases 2 and 3 of the same year.

head(analysis\_df)

## # A tibble: 6 × 25  
## codooii ugel region vl\_2020\_1 vl\_2020\_2 vl\_2020\_2\_3 vl\_2020\_4 vl\_2021\_1\_2  
## <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 10001 UGEL CHA… AMAZO… 0.838 0.37 0.620 NaN 0.725  
## 2 10002 UGEL BAG… AMAZO… 0.923 0.675 0.721 0.969 0.860  
## 3 10003 UGEL BON… AMAZO… 0.919 0.368 0.626 NaN 0.703  
## 4 10004 UGEL CON… AMAZO… 0.767 0.701 0.689 0.986 0.546  
## 5 10005 UGEL LUYA AMAZO… 0.868 0.388 0.639 NaN 0.4   
## 6 10006 UGEL ROD… AMAZO… 0.875 0.368 0.627 NaN 0.746  
## # ℹ 17 more variables: vl\_2021\_3 <dbl>, covid\_deaths\_2020\_1 <dbl>,  
## # covid\_deaths\_2020\_2 <dbl>, covid\_deaths\_2020\_3 <dbl>,  
## # covid\_deaths\_2020\_4 <dbl>, covid\_deaths\_2021\_1 <dbl>,  
## # covid\_deaths\_2021\_2 <dbl>, covid\_deaths\_2021\_3 <dbl>,  
## # covid\_deaths\_2021\_4 <dbl>, covid\_deaths <int>, population <int>,  
## # n\_students <dbl>, rurality <dbl>, mean\_elevation\_msnm <dbl>,  
## # n\_schools <int>, prcnt\_andes <dbl>, prcnt\_selva <dbl>

#First pandemic wave: 2020  
model\_2020 <- lm(vl\_2020\_2\_3 ~ covid\_deaths\_2020\_2 + covid\_deaths\_2020\_3 + n\_students + rurality,   
 data = analysis\_df)  
  
  
#Second pandemic wave: 2021  
model\_2021 <- lm(vl\_2021\_1\_2 ~ covid\_deaths\_2021\_1 + covid\_deaths\_2021\_2 + n\_students + rurality,   
 data = analysis\_df)  
  
#First pandemic hit on performance  
df\_2020\_hit\_full <- analysis\_df %>%   
 mutate(vl\_2020\_dif = vl\_2020\_2\_3 - vl\_2020\_1)  
model\_2020\_hit <- lm(vl\_2020\_dif ~ covid\_deaths\_2020\_2 + covid\_deaths\_2020\_3 + n\_students + rurality,   
 data = df\_2020\_hit\_full)  
  
  
#Models summary  
summary(model\_2020)

##   
## Call:  
## lm(formula = vl\_2020\_2\_3 ~ covid\_deaths\_2020\_2 + covid\_deaths\_2020\_3 +   
## n\_students + rurality, data = analysis\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.195126 -0.027096 0.005527 0.029744 0.228515   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.193e-01 1.341e-02 53.649 < 2e-16 \*\*\*  
## covid\_deaths\_2020\_2 -2.104e-02 6.231e-03 -3.377 0.000868 \*\*\*  
## covid\_deaths\_2020\_3 5.889e-03 5.475e-03 1.076 0.283299   
## n\_students -1.621e-07 7.913e-08 -2.049 0.041661 \*   
## rurality -3.375e-03 1.646e-02 -0.205 0.837773   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05173 on 216 degrees of freedom  
## (2 observations deleted due to missingness)  
## Multiple R-squared: 0.1329, Adjusted R-squared: 0.1169   
## F-statistic: 8.28 on 4 and 216 DF, p-value: 3.127e-06

summary(model\_2021)

##   
## Call:  
## lm(formula = vl\_2021\_1\_2 ~ covid\_deaths\_2021\_1 + covid\_deaths\_2021\_2 +   
## n\_students + rurality, data = analysis\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.46622 -0.08222 0.00823 0.09847 0.32690   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.480e-01 4.463e-02 12.279 <2e-16 \*\*\*  
## covid\_deaths\_2021\_1 5.000e-03 2.508e-02 0.199 0.8422   
## covid\_deaths\_2021\_2 5.030e-02 2.417e-02 2.082 0.0386 \*   
## n\_students -2.109e-07 2.204e-07 -0.957 0.3397   
## rurality 8.873e-02 4.783e-02 1.855 0.0649 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.144 on 218 degrees of freedom  
## Multiple R-squared: 0.0464, Adjusted R-squared: 0.0289   
## F-statistic: 2.652 on 4 and 218 DF, p-value: 0.03413

summary(model\_2020\_hit)

##   
## Call:  
## lm(formula = vl\_2020\_dif ~ covid\_deaths\_2020\_2 + covid\_deaths\_2020\_3 +   
## n\_students + rurality, data = df\_2020\_hit\_full)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.13957 -0.05036 -0.01245 0.03843 0.34862   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.587e-01 1.884e-02 -8.422 5.17e-15 \*\*\*  
## covid\_deaths\_2020\_2 9.090e-03 8.758e-03 1.038 0.3005   
## covid\_deaths\_2020\_3 -4.767e-03 7.696e-03 -0.619 0.5362   
## n\_students -2.100e-07 1.112e-07 -1.889 0.0603 .   
## rurality 1.546e-02 2.314e-02 0.668 0.5047   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.07271 on 216 degrees of freedom  
## (2 observations deleted due to missingness)  
## Multiple R-squared: 0.03684, Adjusted R-squared: 0.019   
## F-statistic: 2.065 on 4 and 216 DF, p-value: 0.08642

# Consclusions and recommendations

In conclusion, our research highlights the significant impact that the first wave of Covid-19 had on UGELs performance in Peru in 2020, particularly in areas that were hit harder by the pandemic. However, our analysis of the second wave in 2021 suggests that institutions may have adapted to the pandemic, and that there may have been underlying factors contributing to performance reduction.

While our findings provide some insights into the effects of Covid-19 on UGELs performance, more research is needed to fully understand the underlying factors contributing to these changes. Further research could include exploring individual indicators evaluated during 2019, 2020, and 2021 to identify trends related to Covid-19 impact on UGELs performance, and incorporating other variables into the linear regression models, such as health access and income in the jurisdiction.

# Exports

export\_sheets <- list("Timeline" = timeline, "Results dataframe"= analysis\_df, "Covid data - UGEL" = covid\_deaths\_ugel)  
openxlsx::write.xlsx(export\_sheets, file = "Analysis Results.xlsx")