DATA 607 Preparing Datasets

David Simbandumwe

2021-09-26

Introduction

The goal of this assignment is to practice preparing different datasets for downstream analysis work.For the purposes of this assignment I choose the following datasets * Global Child Mortality Rates - posted by Alec Mccabe * COVID-19 Mortality Rates in NYC - William Aiken * Annual% GDP Growth - ChunJie Nan

Global Child Mortality Rates - posted by Alec Mccabe

Dataset found here: https://sejdemyr.github.io/r-tutorials/basics/wide-and-long//

This dataset includes child-under-5 mortality rates for all countries from 1950 to 2015. The data is structured in wide format, where the column names include the country, and each year from 1950 to 2015. Values are the corresponding child mortality rates for that country, and that year. Restructuring this dataset into long format should be very easy to accomplish with the tidyr::gather().

This dataset would be a great starting point for analyzing mortality rates for children under 5 over time, by country. It would also be interesting to see if any country mortality rates are correlated over time. Monitoring spikes for mortality rate over time would be a good way to identify patterns or factors leading to child mortality.

Load Data

Loaded the data from a csv file

```
# load data

mortality_df <- read_csv( file = "/Users/dsimbandumwe/dev/cuny/data_607/DATA607/Project2/unicef-u5mr.csv
```

```
## Rows: 196 Columns: 67
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): CountryName
## dbl (66): U5MR 1950, U5MR 1951, U5MR 1952, U5MR 1953, U5MR 1954, U5MR 1955, ...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Tiddy Data

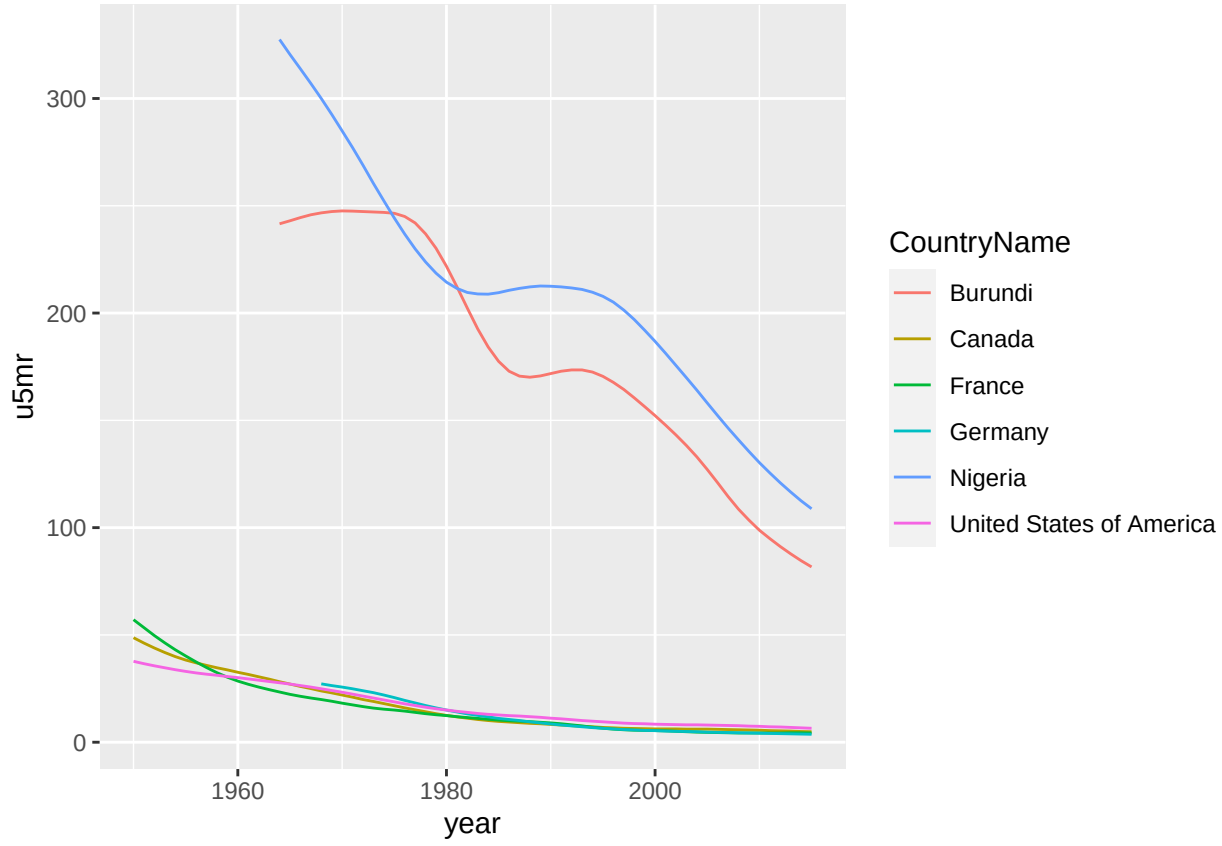
For this analysis i wanted to evaluate the mortality data from three regions. I selected 2 countries from Sub Saharan Africa, Europe and North America.

Steps * Filter the data by selected countries * added the regional flag to the dataset * gathered the year data into rows * updated the year column to numeric * drop all rows with invalid data for u5mr

```
mortality_df <- mortality_df %>%
  filter (CountryName %in% c("Burundi","Nigeria","Germany","France","Canada","United States of America")
  mutate (
    Region = ifelse(CountryName %in% c("Burundi","Nigeria"),"Sub-Saharan Africa",
                    ifelse(CountryName %in% c("Germany","France"),"Europe", "North America")
    )
  ) %>%
  gather(year, u5mr, "U5MR 1950":"U5MR 2015") %>%
  mutate(year = as.numeric(gsub("U5MR.", "", year))) %>%
  drop_na(u5mr)
```

Analysis

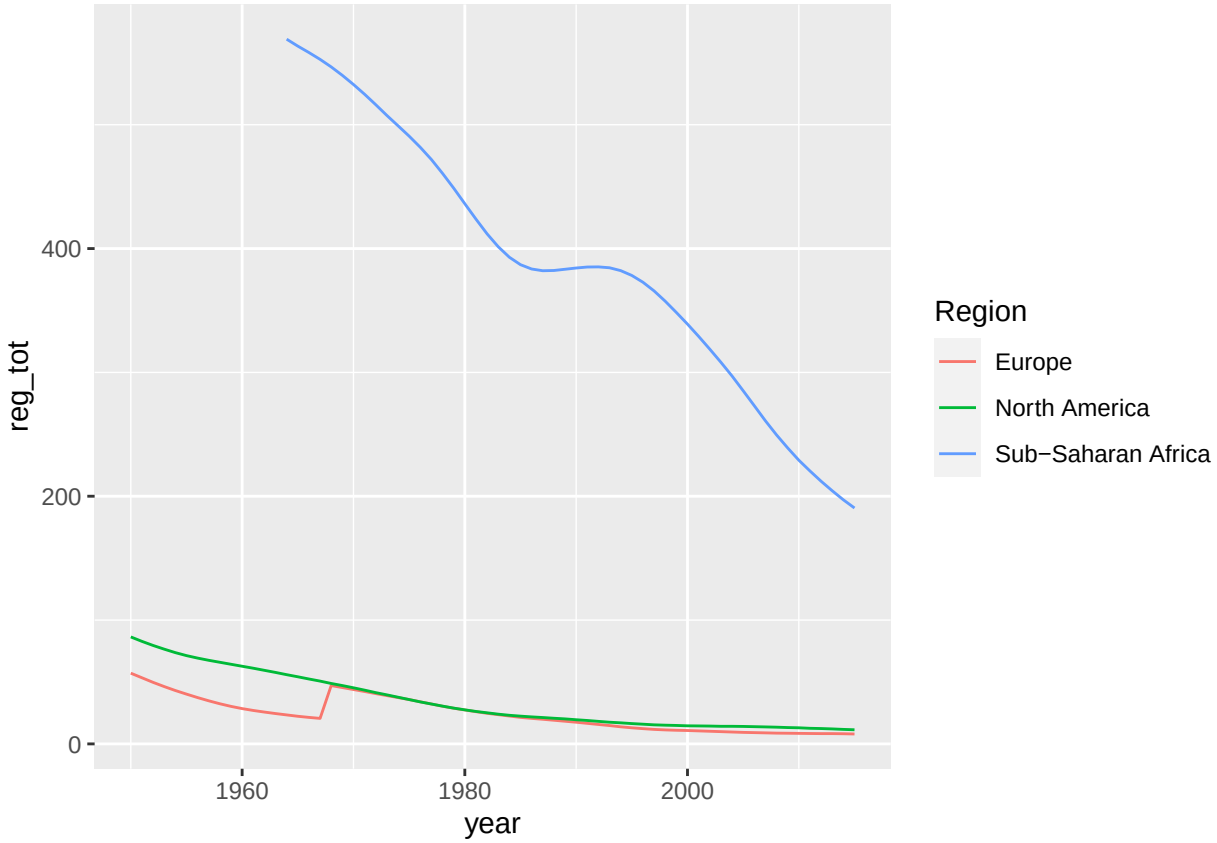
```
ggplot(data = mortality_df) +
  geom_line(mapping = aes(x=year, y=u5mr, color=CountryName))
```



There are large drops in the mortality rates across all countries from 1950 to 2015. However there still

remains a substantial differential in the rates from European and North American countries. The difference between the regions has been shrinking over the years.

```
mortality_df %>%
  group_by(year, Region) %>%
  mutate(
    reg_tot = sum(u5mr)
  ) %>%
  select(Region, year, reg_tot) %>%
  distinct() %>%
  ggplot() +
  geom_line(mapping = aes(x=year, y=reg_tot, color=Region))
```



The regional view of the data shows a similar pattern. All regions are seeing reduced rates of child mortality. With the largest drops in the Sub Saharan Africa region. Interesting to note in the late 60s there was jump in the mortality rate.

#COVID-19 Mortality Rates in NYC The nyc.gov website has both the hopitalization and mortality rates by county and zip code over time. It would be interesting to compare the relationship between these two data set. How it hospitalization and death rates differ by county in NYC?

https://www1.nyc.gov/site/doh/covid/covid-19-data-totals.page

Load Data

Load data from a csv file

```
# load data

covid_df <- read_csv( file = "/Users/dsimbandumwe/dev/cuny/data_607/DATA607/Project2/group-data-by-boro
```

```
## Rows: 18 Columns: 47
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): group, subgroup
## dbl (45): BK_CONFIRMED_CASE_COUNT, BK_PROBABLE_CASE_COUNT, BK_CASE_COUNT, BK...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Tiddy Data

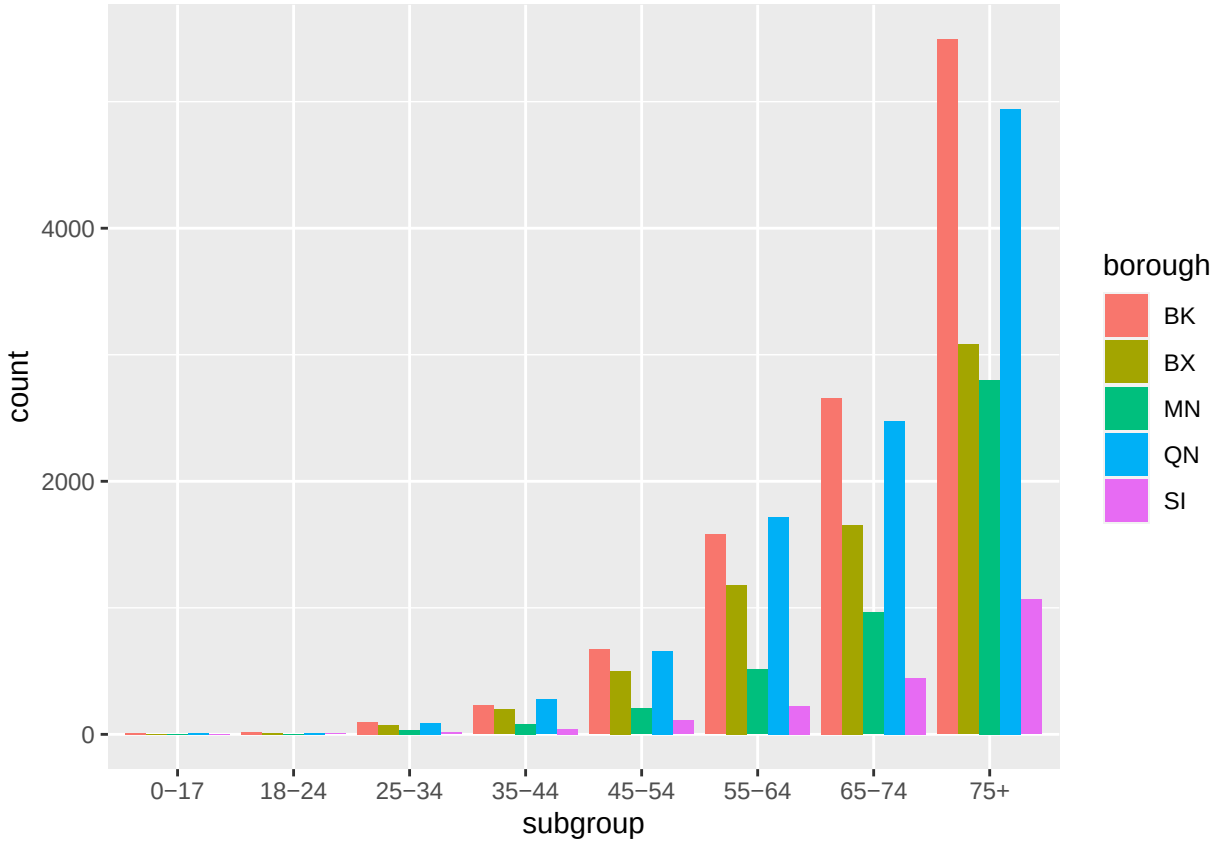
For this analysis I wanted to focus on deaths and hospitalization. Removed the unneeded columns and Transformed the wide data into tall data set.

Steps: * removed the columns that are not needed * gather the columns data to transform the wide dataset into a long dataset * create 2 variable out of the resulting key. one for borough and the other for the type of event * drop the n/a in the count field * clean up the column names

```
covid_df <- covid_df %>%
  select (
    group,
    subgroup,
    BK_HOSPITALIZED_COUNT,
    BK_DEATH_COUNT,
    BX_HOSPITALIZED_COUNT,
    BX_DEATH_COUNT,
    MN_HOSPITALIZED_COUNT,
    MN_DEATH_COUNT,
    QN_HOSPITALIZED_COUNT,
    QN_DEATH_COUNT,
    SI_HOSPITALIZED_COUNT,
    SI_DEATH_COUNT
  ) %>%
  gather(type, count,
    "BK_HOSPITALIZED_COUNT",
    "BK_DEATH_COUNT",
    "BX_HOSPITALIZED_COUNT",
    "BX_DEATH_COUNT",
    "MN_HOSPITALIZED_COUNT",
    "MN_DEATH_COUNT",
    "QN_HOSPITALIZED_COUNT",
    "QN_DEATH_COUNT",
    "SI_HOSPITALIZED_COUNT",
    "SI_DEATH_COUNT"
  ) %>%
  separate(
```

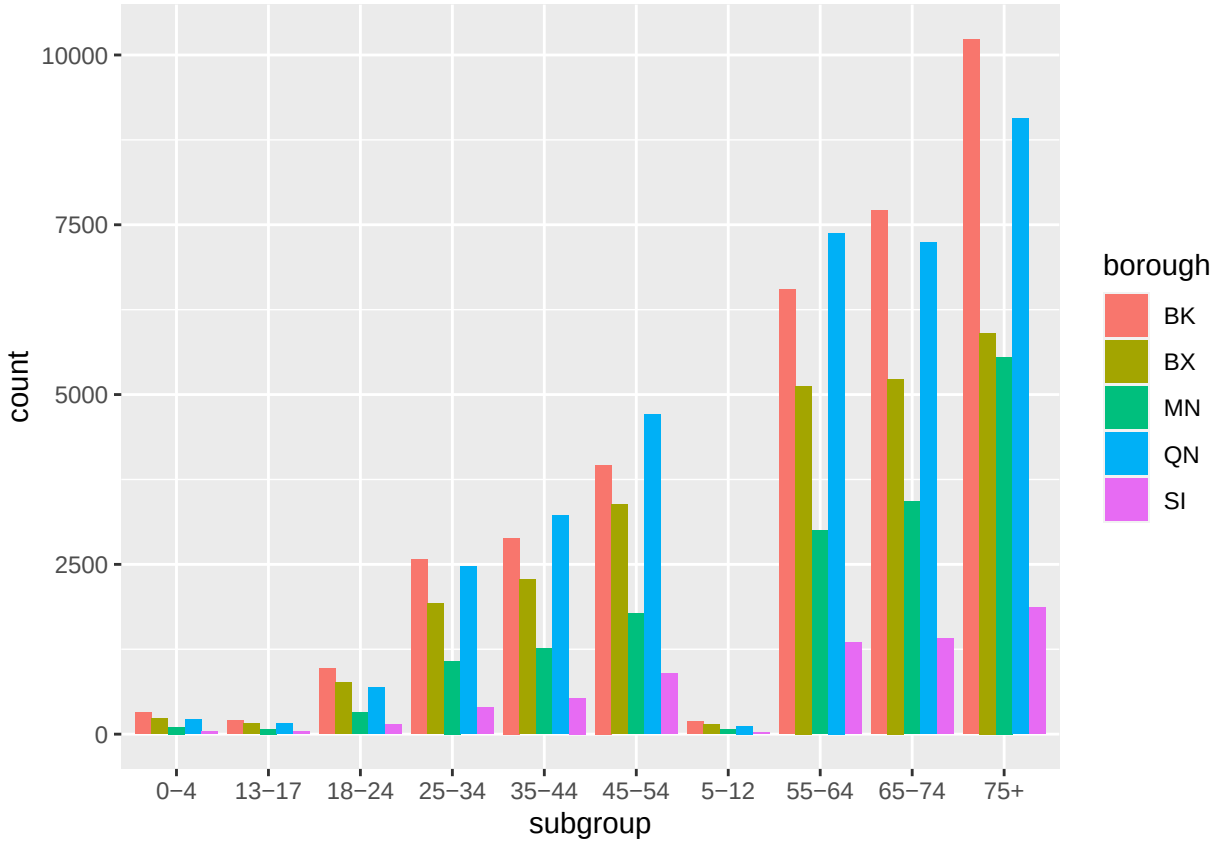
```
    type,
    into = c("borough", "event", "junk"),
    extra = "merge",
    fill = "left",
    convert = TRUE,
    sep = "\\_"
  ) %>%
  drop_na(count)

# drop junk column
covid_df <- covid_df[ , !(names(covid_df) == "junk")]
```

```
covid_df %>%
  filter(group == "Age", event=="DEATH") %>%
  ggplot(aes(fill=borough, y=count, x=subgroup)) +
    geom_bar(position="dodge", stat="identity")
```



The grouped bar chart highlights death rates per age group by borough. As expected the larger boroughs have higher numbers of COVID deaths.

```
covid_df %>%
  filter(group == "Age", event=="HOSPITALIZED") %>%
  ggplot(aes(fill=borough, y=count, x=subgroup)) +
    geom_bar(position="dodge", stat="identity")
```



The grouped bar chart highlights death rates per age group by borough.As expected the larger boroughs have higher numbers of COVID hospitalizations.

##Annual% GDP Growth - ChunJie Nan

The GDP growth(annual%) is a data from The World Bank. It includes 266 observations/countries' % annual GDP growth. The data set is tremendous and has some NA/missing values.

It requires some data adjustment, such as handling the missing value with zoo package and subset the data to a small group especially the top 5 GDP countries to see which high GDP countries has

affected the most by the Covid-19. With the historical data, we can forecast/predict the year 2020 GDP. from the difference between the real value GDP of 2020 and the predicted value of the year

2020 GDP, I can find out how much does Covid-19 affected the GDP growth.

data source:

https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG

Load Data

Load data from a csv file. There is some additional file information at the start of the file so the first line in the file is not a valid list of column names.

```
# load data

gdp_df <- read_csv( file = "/Users/dsimbandumwe/dev/cuny/data_607/DATA607/Project2/gdp.csv",
                    col_names = FALSE,
```

```
                skip_empty_rows = TRUE
                )
```

```
## Rows: 271 Columns: 65
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): X1, X2, X3, X4
## dbl (61): X5, X6, X7, X8, X9, X10, X11, X12, X13, X14, X15, X16, X17, X18, X...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# update column headers
a <- as.list(gdp_df %>% filter(X1 == "Country Name"))
a[1] = "CountryName"
a[2] = "CountryCode"
a[3] = "IndicatorName"
a[4] = "IndicatorCode"
colnames(gdp_df) <- a
```

Tiddy Data

I wanted to analyze regional data so i selected 2 countries from Sub Saharan Africa, Europe and North America respectively. This is a similar dataset from the first part of the assignment so we can combine the two data sets to see if there is any correlation.

The steps taken:filter by the desired countries * add region the country list * gather the wide year data into rows * transform year data into a numeric * join the gdp data with the child mortality data from the first part of the assignment * clean up the rows

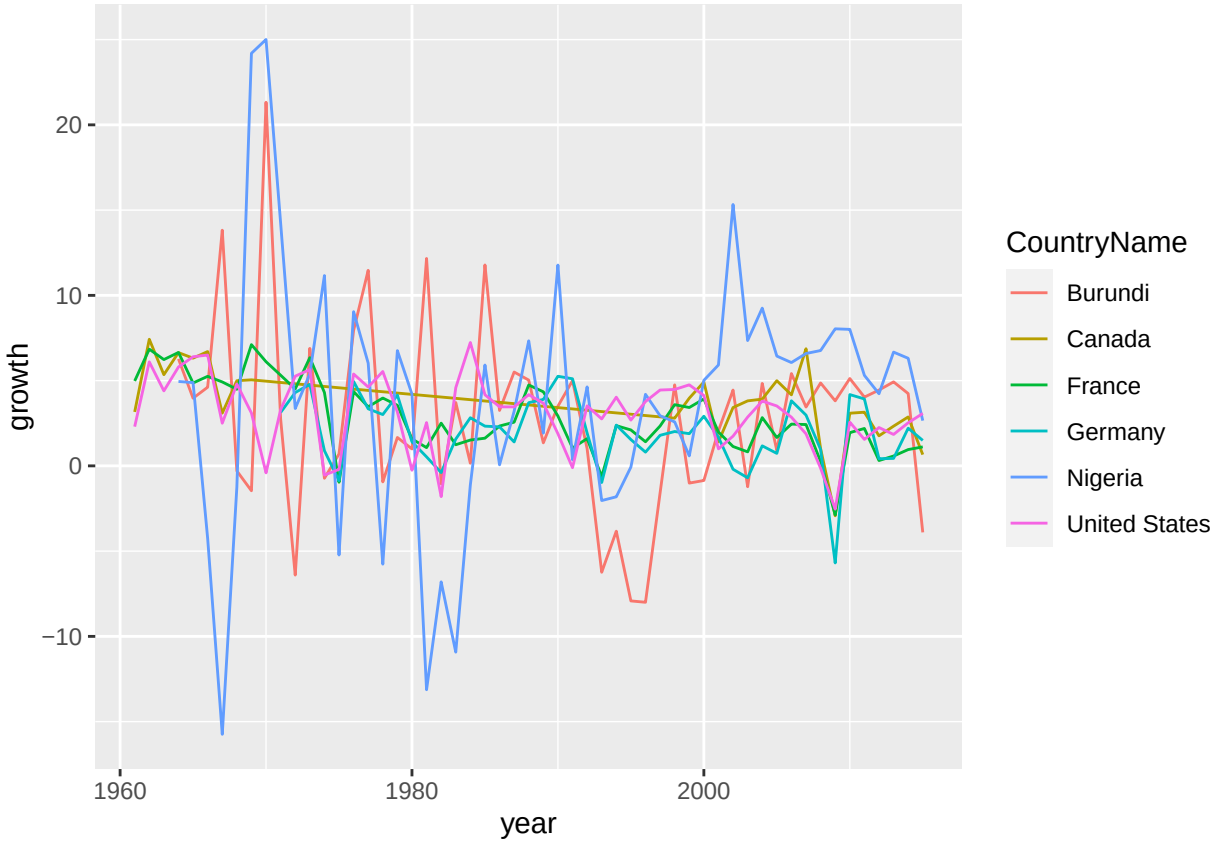
```
# update mortality country field to match gpd data
mortality_df <- mortality_df %>%
  mutate(
    CountryName = ifelse(CountryName == "United States of America","United States", CountryName)
  )
```

```
gdp_df <- gdp_df %>%
  filter (CountryName %in% c("Burundi","Nigeria","Germany","France","Canada","United States")) %>%
  mutate (
    Region = ifelse(CountryName %in% c("Burundi","Nigeria"),"Sub-Saharan Africa",
                    ifelse(CountryName %in% c("Germany","France"),"Europe", "North America")
    )
  ) %>%
  gather(year, growth, "1960":"2020") %>%
  drop_na(growth) %>%
  select(CountryName, CountryCode,Region,year,growth) %>%
  transform(year = as.numeric(year))
```

```
comb_df <- gdp_df %>%
  left_join(mortality_df, by = c("CountryName" = "CountryName","year" = "year" )) %>%
  drop_na(u5mr) %>%
  rename(
      region =  Region.x
  )

  comb_df <- comb_df[ , !(names(comb_df) == "Region.y")]
```

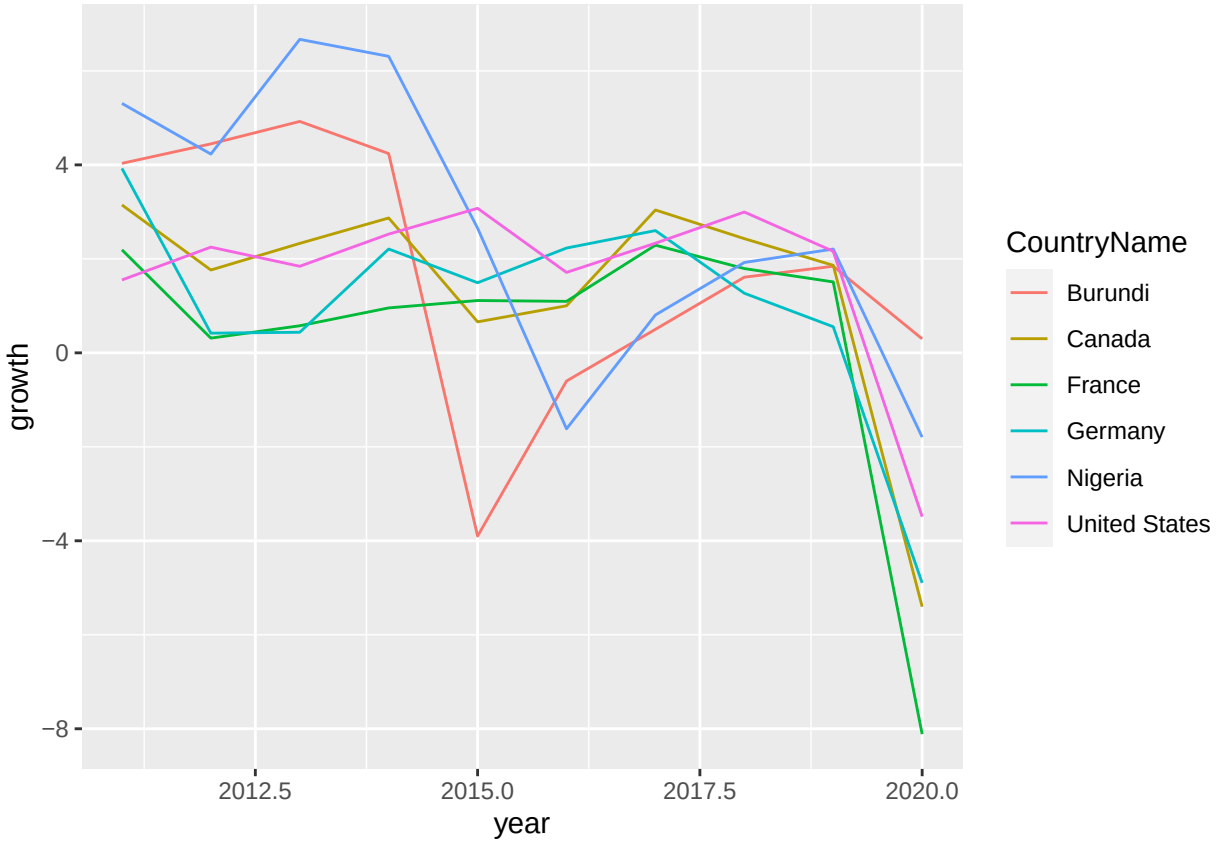
Analysis

```
ggplot(data = comb_df) +
  geom_line(mapping = aes(x=year, y=growth, color=CountryName))
```



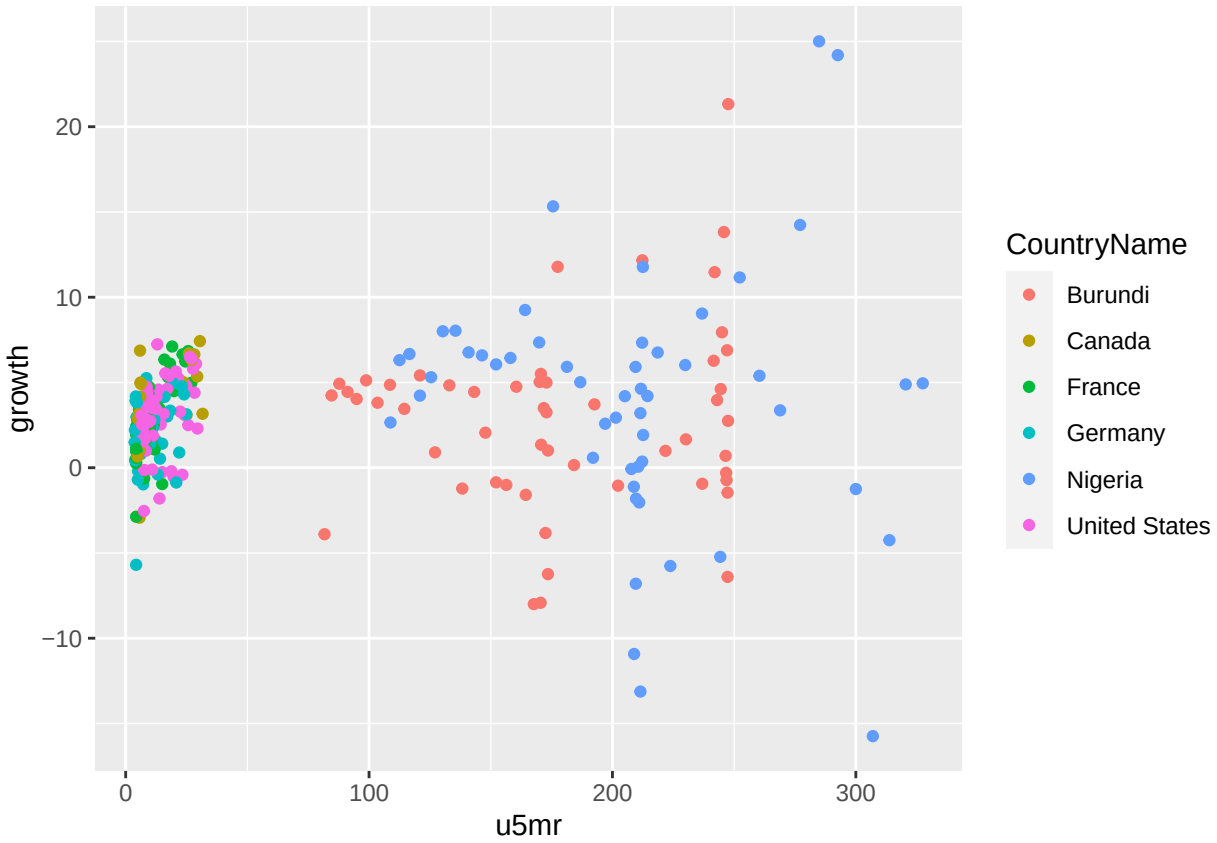
The growth in gdp by country over the years shows that countries in Europe and North America see more stability in their year over year growth numbers. While countries in Sub-Saharan Africa see a larger fluctuation in GDP.

```
gdp_df %>%
  filter(year > 2010) %>%
    ggplot() +
    geom_line(mapping = aes(x=year, y=growth, color=CountryName))
```



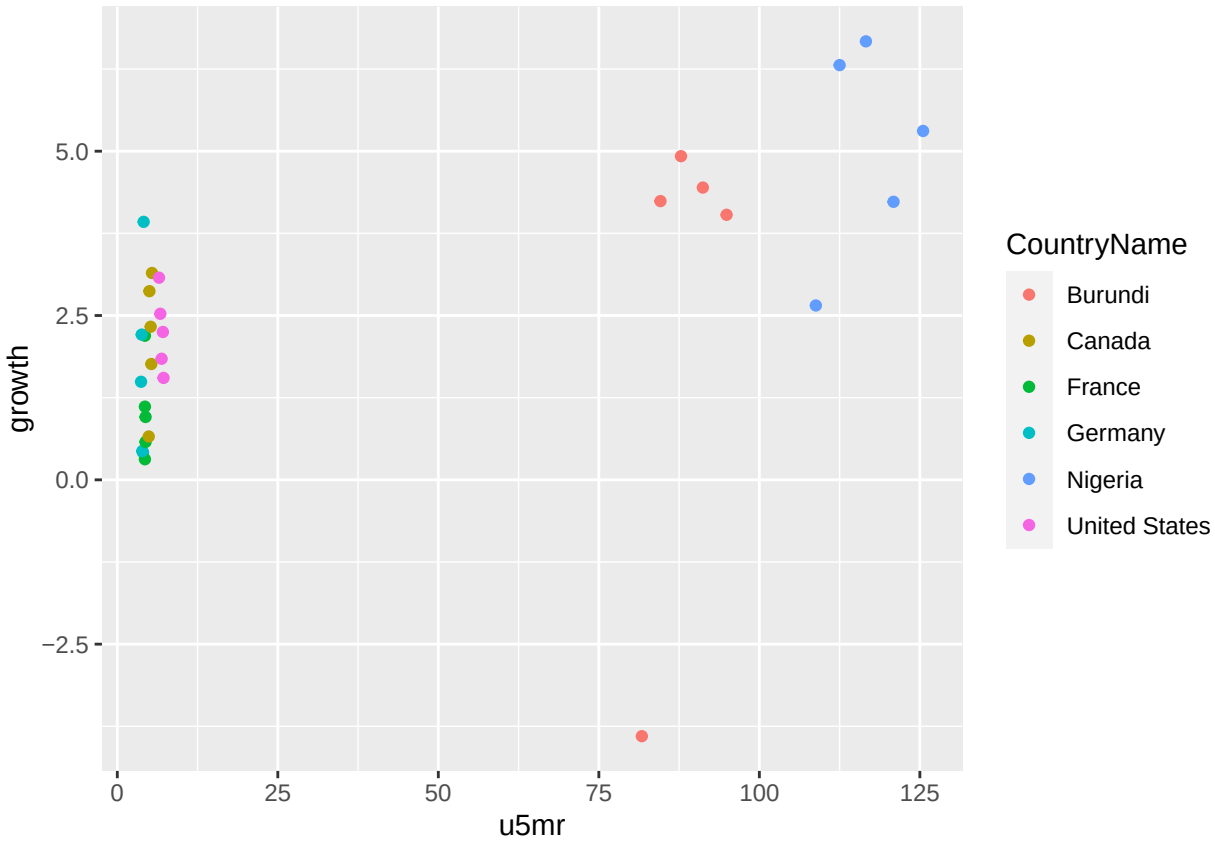
If we zoom in to 2020 we can see the impact of COVID on economies across the world. It is interesting that COVID impacted countries in Europ and North America more than it impacted countries in Sub-Saharan Africa.

```
ggplot(data = comb_df) +
  geom_point(mapping = aes(x=u5mr, y=growth, color=CountryName))
```

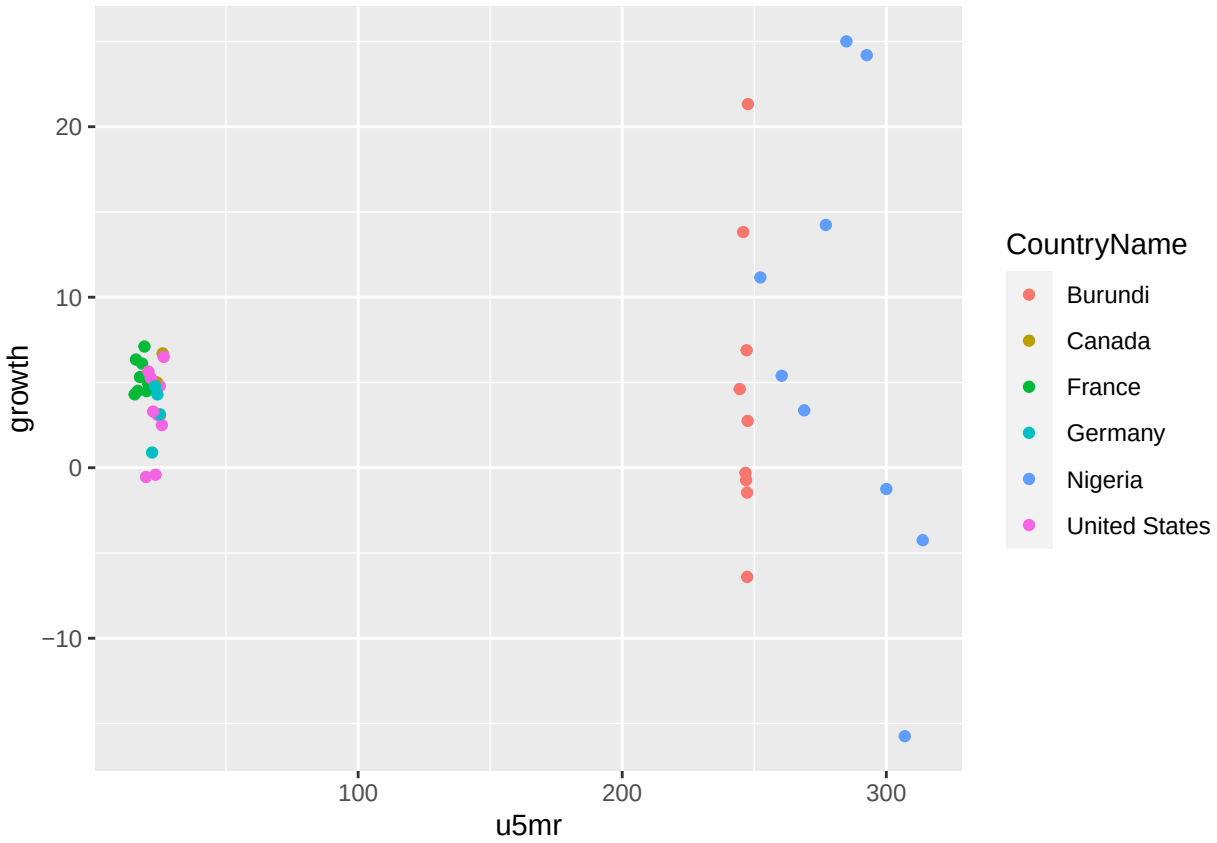


When reviewing GDP growth mapped against child mortality the data for Europe and North America are tightly clustered around 0% growth and low child mortality. The data for Sub-Saharan Africa is a little mixed.

```
comb_df %>%
  filter(year > 2010) %>%
  ggplot() +
  geom_point(mapping = aes(x=u5mr, y=growth, color=CountryName))
```




```
comb_df %>%
  filter(year > 1965, year < 1975) %>%
  ggplot() +
  geom_point(mapping = aes(x=u5mr, y=growth, color=CountryName))
```



The results from Burundi and Nigeria during the 1965 to 1975 timeframe is very interesting. This was the point in time when the mortality rates in Nigeria were dropping faster than the rates in Burundi. The gdp growth rates were almost identical during that timeframe pointing to the potential for another source for the reductions.