Global Covid-19 Analysis

Rushti Rajoli

23/12/2021

Loading all installed packages

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

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

```
library(skimr)
library(readxl)
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

importing and previewing Covid 19 Dataset

```
covid_death <- read_excel("F:/Covid Project/covid_death.xlsx")
skim_without_charts(covid_death)
```

Table 1: Data summary

Name	covid_death
Number of rows	148790
Number of columns	26
Column type frequency:	
character	3
logical	8
numeric	14
POSIXct	1
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
iso_code	0	1.00	3	8	0	238	0
continent	8981	0.94	4	13	0	6	0
location	0	1.00	4	32	0	238	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
icu_patients	131488	0.12	0.98	TRU: 16875, FAL: 427
icu_patients_per_million	131488	0.12	0.92	TRU: 15951, FAL: 1351
hosp_patients	129032	0.13	0.99	TRU: 19551, FAL: 207
hosp_patients_per_million	129032	0.13	0.98	TRU: 19449, FAL: 309
weekly_icu_admissions	147443	0.01	0.94	TRU: 1262, FAL: 85
weekly_icu_admissions_per_million	147443	0.01	0.84	TRU: 1132, FAL: 215
weekly_hosp_admissions	146549	0.02	0.98	TRU: 2202, FAL: 39
weekly_hosp_admissions_per_million	146549	0.02	0.97	TRU: 2177, FAL: 64

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
population	1000	0.99	1487499687[illegible]	[illegible]87743244[illegible]	47.00	1273428.00	8715494.00	3393361170.00	7.874966e+09
total_cases	2600	0.98	2046805.70	12151675.29	1.00	1462.00	19220.00	244010.25	2.754609e+08
new_cases	2606	0.98	7876.76	42989.50	-74347.00	1.00	67.00	898.00	9.082890e+05
new_cases_smoothed	3751	0.97	7881.07	42361.01	-6223.00	5.57	87.57	949.00	8.272200e+05
total_deaths	19590	0.87	51197.62	269627.44	1.00	65.00	645.00	6262.00	5.360875e+06
new_deaths	19394	0.87	172.03	838.73	-1918.00	0.00	2.00	19.00	1.800700e+04
new_deaths_smoothed	3751	0.97	152.93	775.27	-232.14	0.00	1.29	15.00	1.470314e+04
total_cases_per_million	3283	0.98	22245.94	36389.59	0.00	480.78	3849.35	28682.65	2.722807e+05

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
new_cases_per_milli	32289	0.98	96.22	298.19	-3125.83	0.02	9.64	83.31	5.142749e+04
new_cases_smoothed_per_million	44290	0.97	96.06	214.70	-272.97	1.39	15.23	95.76	7.406210e+03
total_deaths_per_mi	201260	0.86	440.57	702.52	0.00	15.04	100.58	591.74	6.062010e+03
new_deaths_per_mi	200664	0.87	1.63	5.08	-75.91	0.00	0.12	1.28	4.537700e+02
new_deaths_smoothed_per_million	44290	0.97	1.45	3.49	-10.84	0.00	0.16	1.30	1.441700e+02
reproduction_rate	36780	0.75	1.00	0.34	-0.03	0.83	1.00	1.17	5.920000e+00

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2020-01-01	2021-12-20	2021-02-03	720

Selecting and Sorting Dataset

selecting required columns and and sorted by location and date in ascending order and stored in new variable named death_data

```
covid_death %>%
 select(continent,location, date,total_cases,new_cases,total_deaths,new_deaths,population) %>% arrange
```

```
## # A tibble: 148,790 x 8
##    continent location    date                total_cases new_cases total_deaths
##    <chr>     <chr>       <dttm>                    <dbl>     <dbl>        <dbl>
##  1 Asia      Afghanistan 2020-02-24 00:00:00           5         5           NA
##  2 Asia      Afghanistan 2020-02-25 00:00:00           5         0           NA
##  3 Asia      Afghanistan 2020-02-26 00:00:00           5         0           NA
##  4 Asia      Afghanistan 2020-02-27 00:00:00           5         0           NA
##  5 Asia      Afghanistan 2020-02-28 00:00:00           5         0           NA
##  6 Asia      Afghanistan 2020-02-29 00:00:00           5         0           NA
##  7 Asia      Afghanistan 2020-03-01 00:00:00           5         0           NA
##  8 Asia      Afghanistan 2020-03-02 00:00:00           5         0           NA
##  9 Asia      Afghanistan 2020-03-03 00:00:00           5         0           NA
## 10 Asia      Afghanistan 2020-03-04 00:00:00           5         0           NA
## # ... with 148,780 more rows, and 2 more variables: new_deaths <dbl>,
## #   population <dbl>
```

Cleaning dataset

removing NULL values from every columns and filtering 2020 data alone

```
 covid_death %>%
 select(continent,location, date,total_cases,new_cases,total_deaths,new_deaths,population) %>% arrange
 drop_na() %>%
 filter(date >= "2020-01-01" & date <= "2020-12-31")
```

```
## # A tibble: 50,513 x 8
##    continent location    date                total_cases new_cases total_deaths
##    <chr>     <chr>       <dttm>                    <dbl>     <dbl>        <dbl>
##  1 Asia      Afghanistan 2020-03-23 00:00:00          40         6            1
##  2 Asia      Afghanistan 2020-03-24 00:00:00          42         2            1
##  3 Asia      Afghanistan 2020-03-25 00:00:00          74        32            1
##  4 Asia      Afghanistan 2020-03-26 00:00:00          80         6            2
##  5 Asia      Afghanistan 2020-03-27 00:00:00          91        11            2
##  6 Asia      Afghanistan 2020-03-28 00:00:00         106        15            2
##  7 Asia      Afghanistan 2020-03-29 00:00:00         114         8            4
##  8 Asia      Afghanistan 2020-03-30 00:00:00         114         0            4
##  9 Asia      Afghanistan 2020-03-31 00:00:00         166        52            4
## 10 Asia      Afghanistan 2020-04-01 00:00:00         192        26            4
## # ... with 50,503 more rows, and 2 more variables: new_deaths <dbl>,
## #   population <dbl>
```

Exploring dataset

looking at (total death and total cases vs population) added new columns

```
covid_death %>%
 select(continent,location, date, population, total_cases, new_cases, total_deaths, new_deaths)  %>%
 drop_na()  %>%
 filter(date >= "2020-01-01" & date <= "2020-12-31") %>%
 mutate(total_case_percent = (total_cases/population)*100,
        total_death_percent = (total_deaths/population)*100)
```

```
## # A tibble: 50,513 x 10
##    continent location    date                population total_cases new_cases
##    <chr>     <chr>       <dttm>                   <dbl>       <dbl>     <dbl>
##  1 Asia      Afghanistan 2020-03-23 00:00:00   39835428          40         6
##  2 Asia      Afghanistan 2020-03-24 00:00:00   39835428          42         2
##  3 Asia      Afghanistan 2020-03-25 00:00:00   39835428          74        32
##  4 Asia      Afghanistan 2020-03-26 00:00:00   39835428          80         6
##  5 Asia      Afghanistan 2020-03-27 00:00:00   39835428          91        11
##  6 Asia      Afghanistan 2020-03-28 00:00:00   39835428         106        15
##  7 Asia      Afghanistan 2020-03-29 00:00:00   39835428         114         8
##  8 Asia      Afghanistan 2020-03-30 00:00:00   39835428         114         0
##  9 Asia      Afghanistan 2020-03-31 00:00:00   39835428         166        52
## 10 Asia      Afghanistan 2020-04-01 00:00:00   39835428         192        26
## # ... with 50,503 more rows, and 4 more variables: total_deaths <dbl>,
## #   new_deaths <dbl>, total_case_percent <dbl>, total_death_percent <dbl>
```

Aggregating dataset

Creating, (total cases and total deaths) column to see total cases and total deaths respectively in 2020 and grouped by continent and countries

```
covid_death %>%
 select(continent,location, date, population, total_cases, new_cases, total_deaths, new_deaths)  %>%
 drop_na()  %>%
 filter(date >= "2020-01-01" & date <= "2020-12-31") %>%
```

```
  mutate(total_case_percent = (total_cases/population)*100,
         total_death_percent = (total_deaths/population)*100) %>%
  group_by(continent,location) %>%
  summarise(total_cases_2020 = sum(new_cases), total_deaths_2020 = sum(new_deaths))
```

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

```
## # A tibble: 187 x 4
## # Groups:   continent [6]
##    continent location                 total_cases_2020 total_deaths_2020
##    <chr>     <chr>                               <dbl>             <dbl>
##  1 Africa    Algeria                             99291              2751
##  2 Africa    Angola                              17428               405
##  3 Africa    Benin                                3229                44
##  4 Africa    Botswana                            14697                40
##  5 Africa    Burkina Faso                         6616                84
##  6 Africa    Burundi                               804                 2
##  7 Africa    Cameroon                            26211               448
##  8 Africa    Cape Verde                          11790               112
##  9 Africa    Central African Republic             4484                63
## 10 Africa    Chad                                 2031               104
## # ... with 177 more rows
```

Visualizing dataset

Creating visualization for total cases in 2020 and total deaths in 2020 by continents

```
covid_death %>%
  select(continent,location, date, population, total_cases, new_cases, total_deaths, new_deaths)  %>%
  drop_na()  %>%
  filter(date >= "2020-01-01" & date <= "2020-12-31") %>%
  mutate(total_case_percent = (total_cases/population)*100,
         total_death_percent = (total_deaths/population)*100) %>%
  group_by(continent) %>%
  summarise(total_cases_2020 = sum(new_cases), total_deaths_2020 = sum(new_deaths))
```

```
## # A tibble: 6 x 3
##   continent     total_cases_2020 total_deaths_2020
##   <chr>                    <dbl>             <dbl>
## 1 Africa                 2723869             64768
## 2 Asia                  19798551            336399
## 3 Europe                23614036            539898
## 4 North America         22862962            508259
## 5 Oceania                  46824              1059
## 6 South America         13104440            412389
```

```
covid_death %>%
  arrange(location,date) %>%
  select(continent,location, date, population, total_cases, new_cases, total_deaths, new_deaths)  %>%
  drop_na() %>%
  filter(date >= "2020-01-01" & date <= "2020-12-31") %>%
```

```
  mutate(total_case_percent = (total_cases/population), total_death_percent = (total_deaths/population)
 group_by(continent,location)  %>%
 summarise(total_cases_2020 = sum(new_cases), total_deaths_2020 = sum(new_deaths))  %>%
 ggplot()+geom_smooth(mapping = aes(x= total_cases_2020, y= total_deaths_2020, color = continent ))+
 scale_x_continuous(labels = comma)+
 scale_y_continuous(labels = comma)+
 labs(title = "Covid 19", subtitle = "Total Cases vs Total Deaths", caption = "Viz by Rushti Rajoli")
```

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

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at -119.79
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1830.8
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 7.1882e+008
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : span too small. fewer
## data values than degrees of freedom.
```

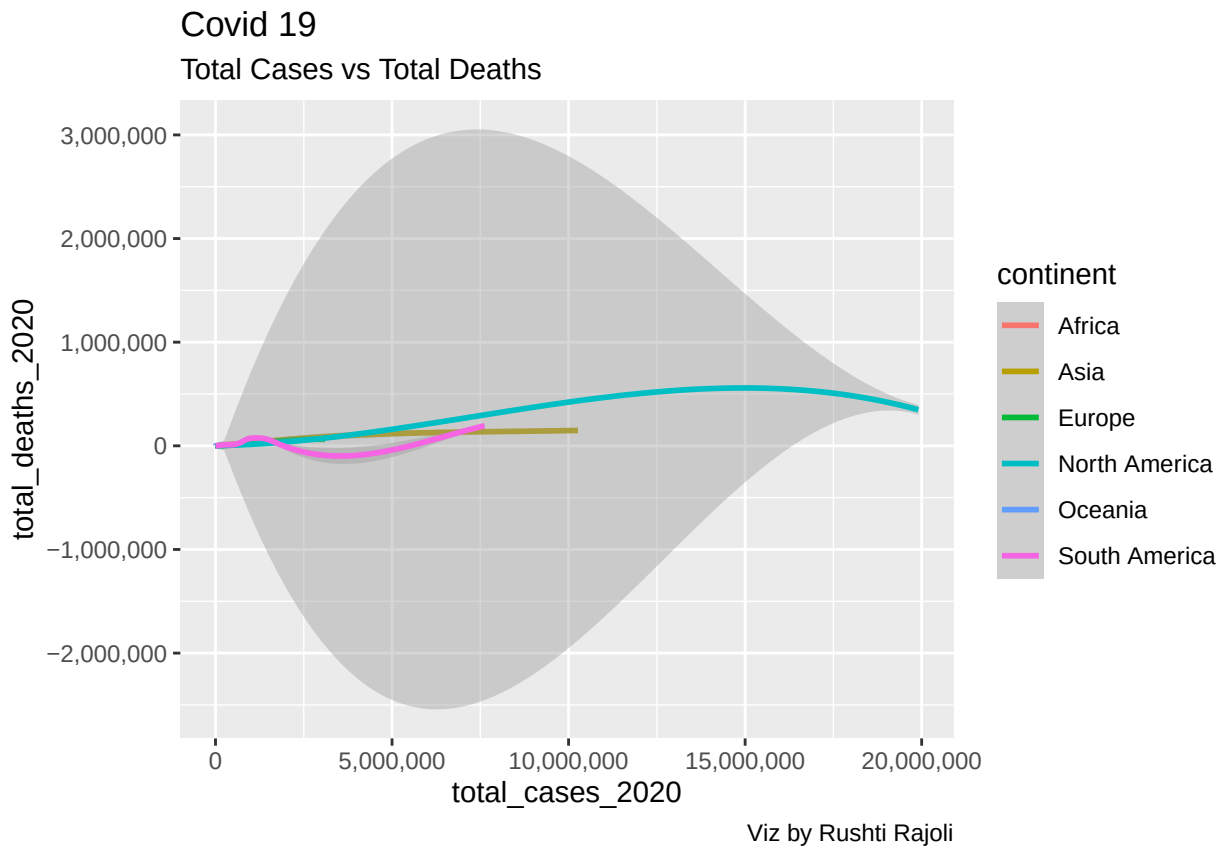
```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## -119.79
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 1830.8
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 7.1882e+008
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```



Verifying the findings

verifying dataset to see overall "India" data (30th january 2020 - 21st December 2021)

```
covid_death %>% arrange(location,date) %>%
  select(continent,location, date, population, total_cases, new_cases, total_deaths, new_deaths)  %>%
  drop_na() %>%
  mutate(total_case_percent = (total_cases/population), total_death_percent = (total_deaths/population))
  group_by(continent,location)  %>% summarise(total_cases = sum(new_cases), total_deaths = sum(new_deat
```

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

```
## # A tibble: 1 x 4
## # Groups:   continent [1]
##   continent location total_cases total_deaths
##   <chr>     <chr>          <dbl>        <dbl>
## 1 Asia      India       34746782       477554
```