Global Covid Data Repository

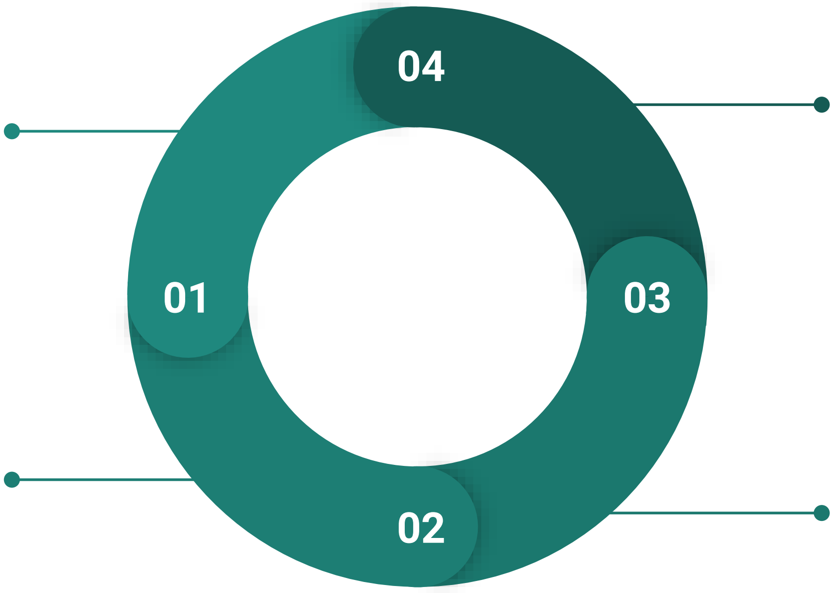
DATA 422 Group Project

Emily, Sarmilan, Ajay, Wuqiu, Thanh

About the Dataset

- Wikipedia articles and UN statistical database
- 219 countries
- Variables : GDP, Population Density, MtoF ratio etc
- Can help Governments with policy reforms

Process



Scraping

Scrape and consolidate covid data from Wikipedia articles and UN website.

Wrangling

Perform data validation, cleaning and transformation to ensure accuracy and reliability.

Analysis

Create visualisations to highlight key trends in the data.

Data Modelling

Create a data model to establish relationship between the different tables.

Project Workflow

Github - Version Control

- To track code changes and roll back to previous versions seamlessly.
- Fosters collaboration with team.
- Can be used to perform code review.

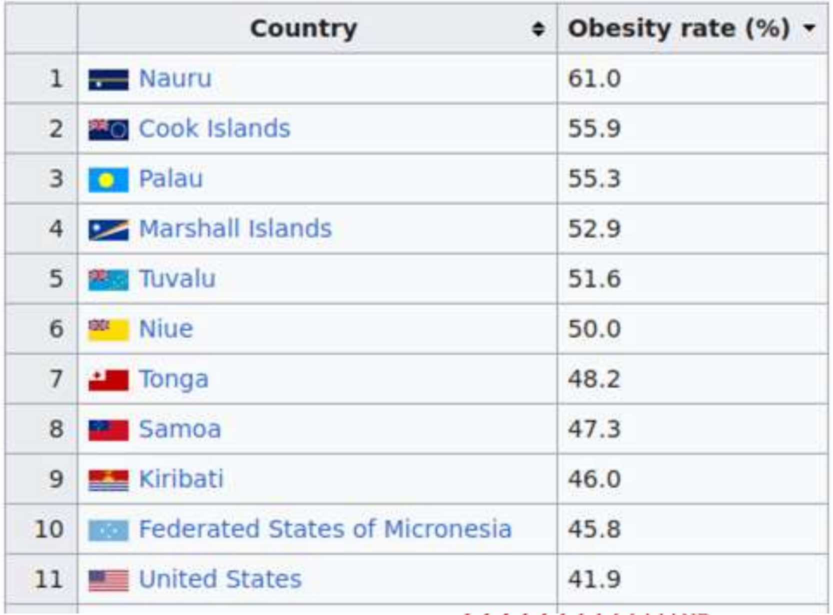
Trello - Project Management

- Provides a visual representation of tasks, their status, and who's working on them.
- Allows task prioritisation.
- Empowers the team by adopting the ‘pull approach’ to task assignment so members can work at their own pace flexibly.

Data Scraping - Wiki Tables

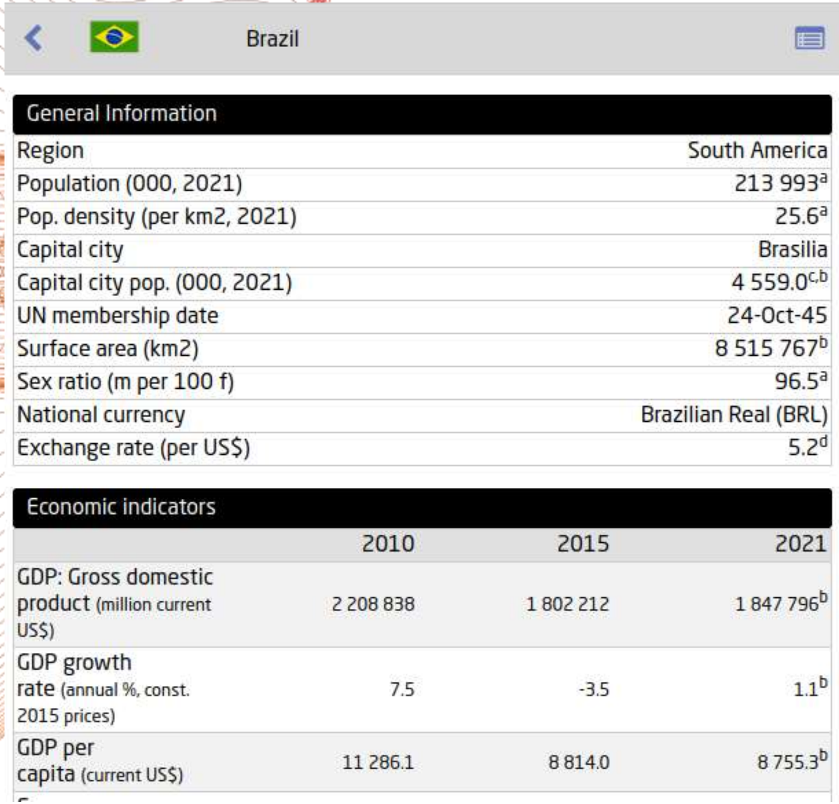
Three tables, consistent country names

Pre-cleaning footnotes & symbols



	Country	Obesity rate (%)
1	Nauru	61.0
2	Cook Islands	55.9
3	Palau	55.3
4	Marshall Islands	52.9
5	Tuvalu	51.6
6	Niue	50.0
7	Tonga	48.2
8	Samoa	47.3
9	Kiribati	46.0
10	Federated States of Micronesia	45.8
11	United States	41.9

Data Scraping - UNdata



Brazil

General Information	
Region	South America
Population (000, 2021)	213 993[a]
Pop. density (per km2, 2021)	25.6[a]
Capital city	Brasilia
Capital city pop. (000, 2021)	4 559.0[c,b]
UN membership date	24-Oct-45
Surface area (km2)	8 515 767[b]
Sex ratio (m per 100 f)	96.5[a]
National currency	Brazilian Real (BRL)
Exchange rate (per US$)	5.2[d]

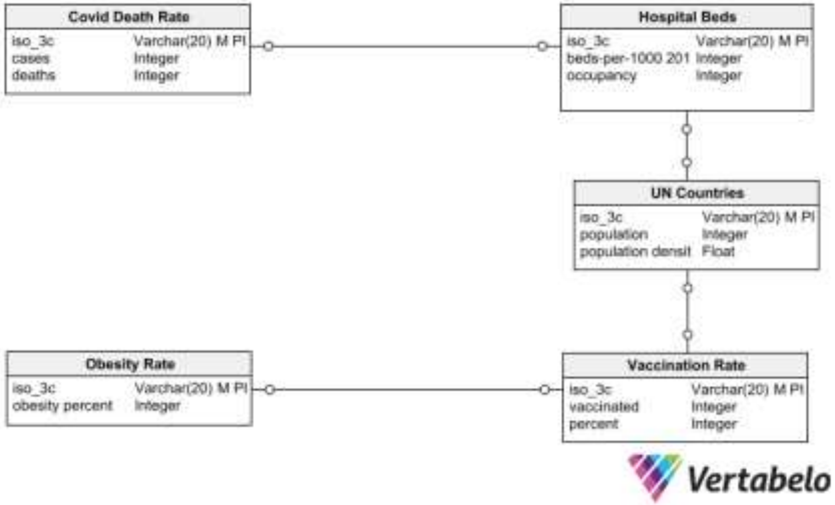
Economic indicators			
	2010	2015	2021
GDP: Gross domestic product (million current US$)	2 208 838	1 802 212	1 847 796[b]
GDP growth rate (annual %, const. 2015 prices)	7.5	-3.5	1.1[b]
GDP per capita (current US$)	11 286.1	8 814.0	8 755.3[b]

01 List of countries and local links

02 Three to four tables per country

Data Model

- Used the CountryCode package to convert country names into iso3 codes that can act as the primary key.



Covid Death Rate	
iso_3c	Varchar(20) M PI
cases	Integer
deaths	Integer

Hospital Beds	
iso_3c	Varchar(20) M PI
beds-per-1000 201	Integer
occupancy	Integer

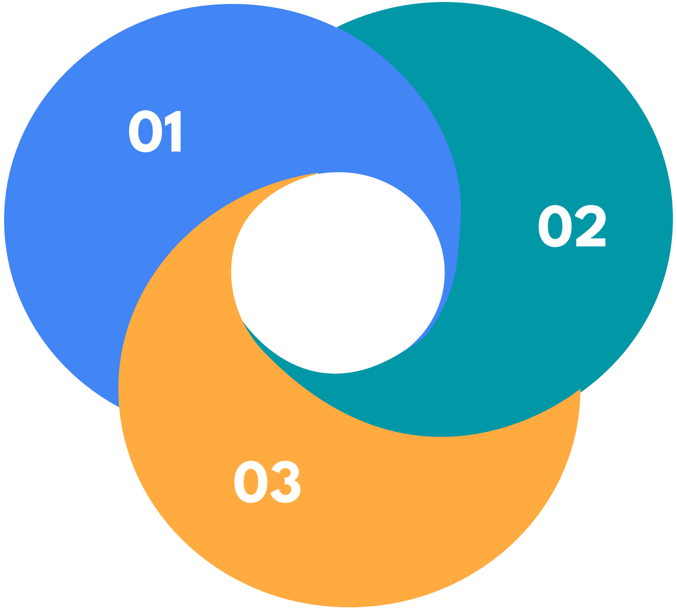
UN Countries	
iso_3c	Varchar(20) M PI
population	Integer
population densit	Float

Obesity Rate	
iso_3c	Varchar(20) M PI
obesity percent	Integer

Vaccination Rate	
iso_3c	Varchar(20) M PI
vaccinated	Integer
percent	Integer

Data Wrangling

Handling Missing Values



Ensuring Unique Identifiers

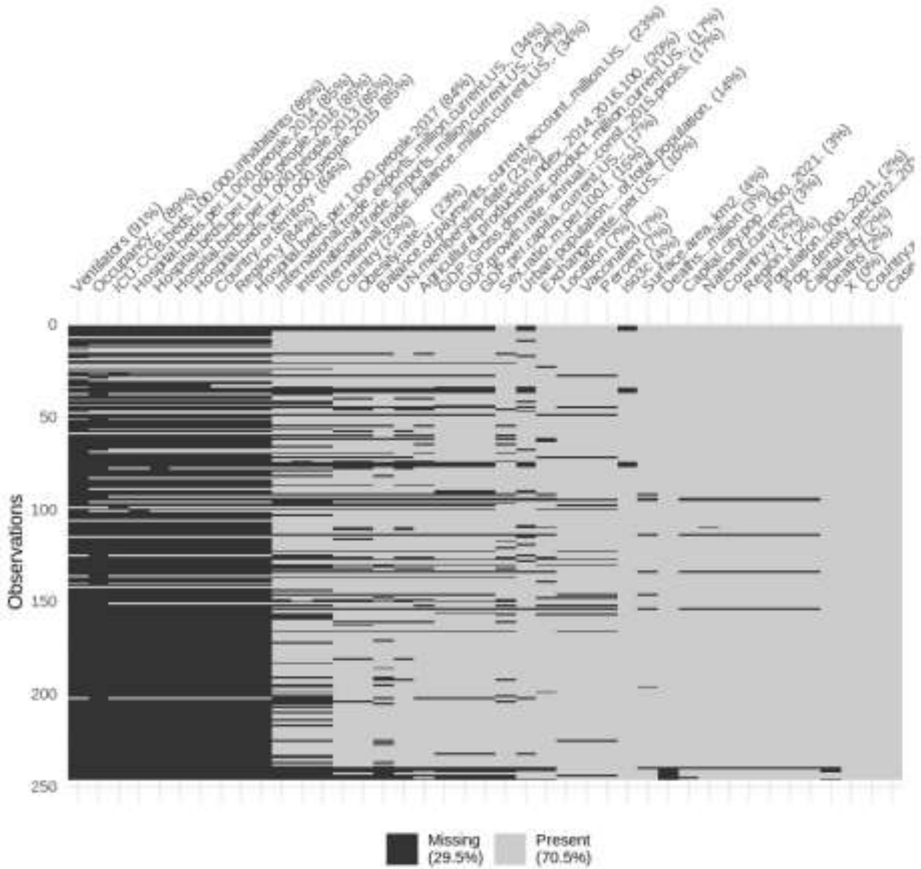
Data Transformation

Handling Missing Values

The vis_miss function from the visdat package is used to visualise the missingness of each variable in the dataset.

The percentage of missing values for each variable is calculated.

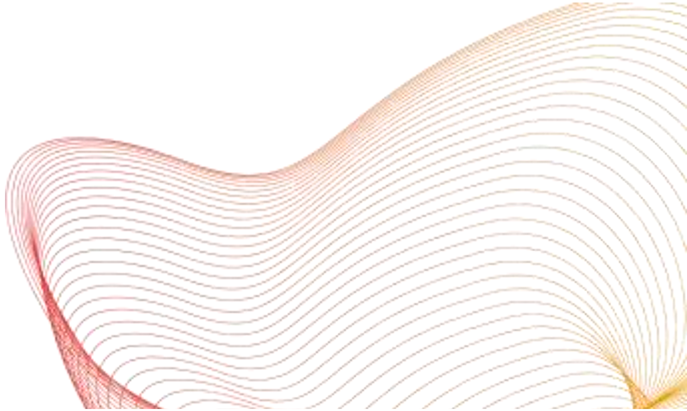
Variables with less than 50% missing values are retained, while others are discarded.

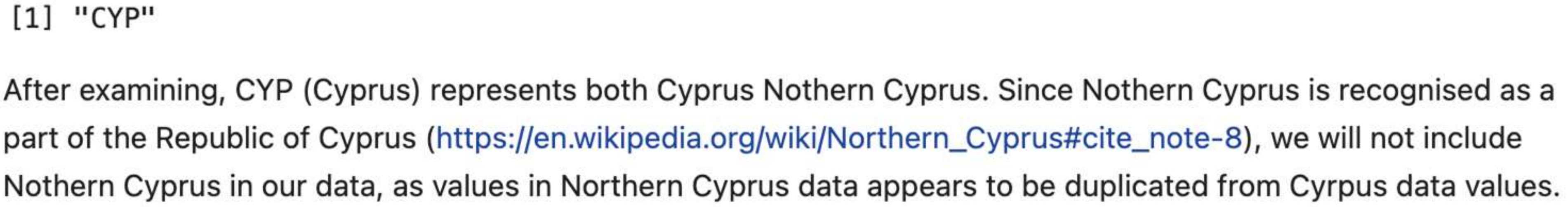


Ensuring Unique Identifiers

- Rows with NA values in the iso3c column are removed, as iso3c will be used as unique identifier for each country.

- Duplicated rows based on the iso3c column are identified and displayed.

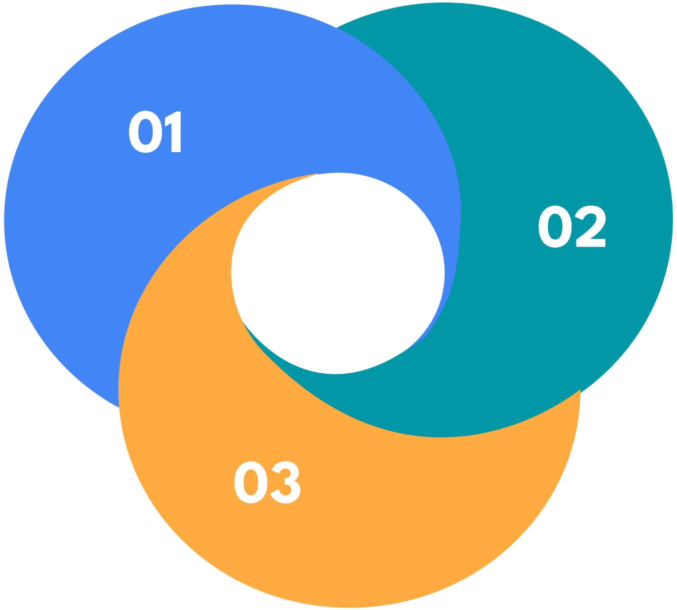




```
[1] "CYP"
```

After examining, CYP (Cyprus) represents both Cyprus Nothern Cyprus. Since Nothern Cyprus is recognised as a part of the Republic of Cyprus (https://en.wikipedia.org/wiki/Northern_Cyprus#cite_note-8), we will not include Nothern Cyprus in our data, as values in Northern Cyprus data appears to be duplicated from Cyrpus data values.

Data Transformation

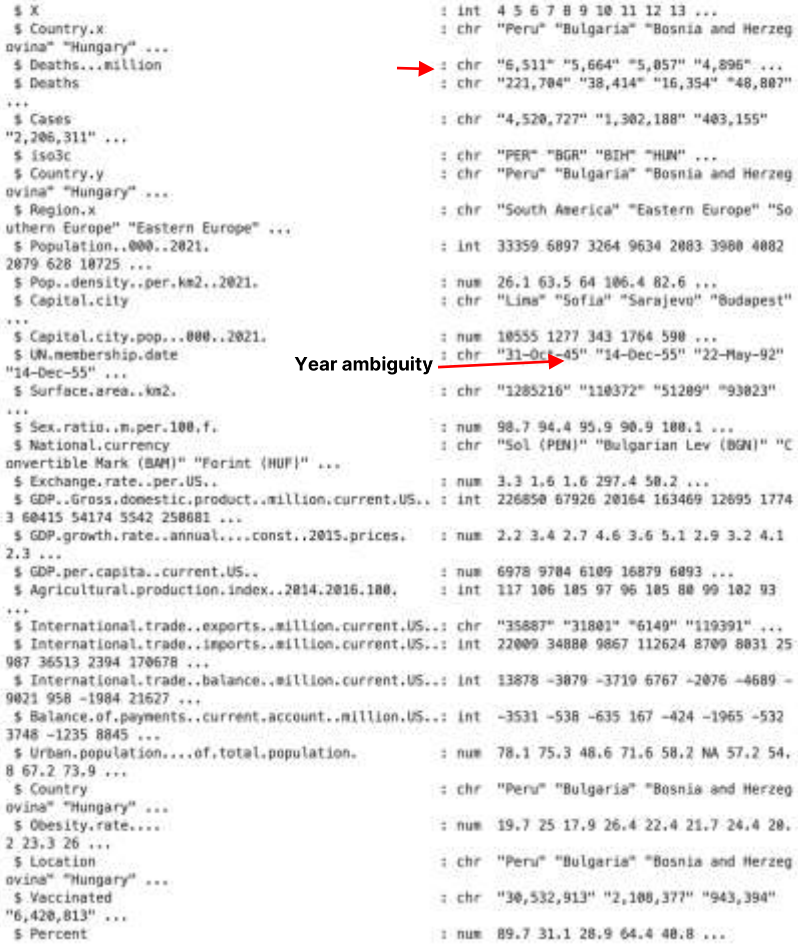


01 Convert numerical columns to appropriate data type

02 Remove redundant location related columns.

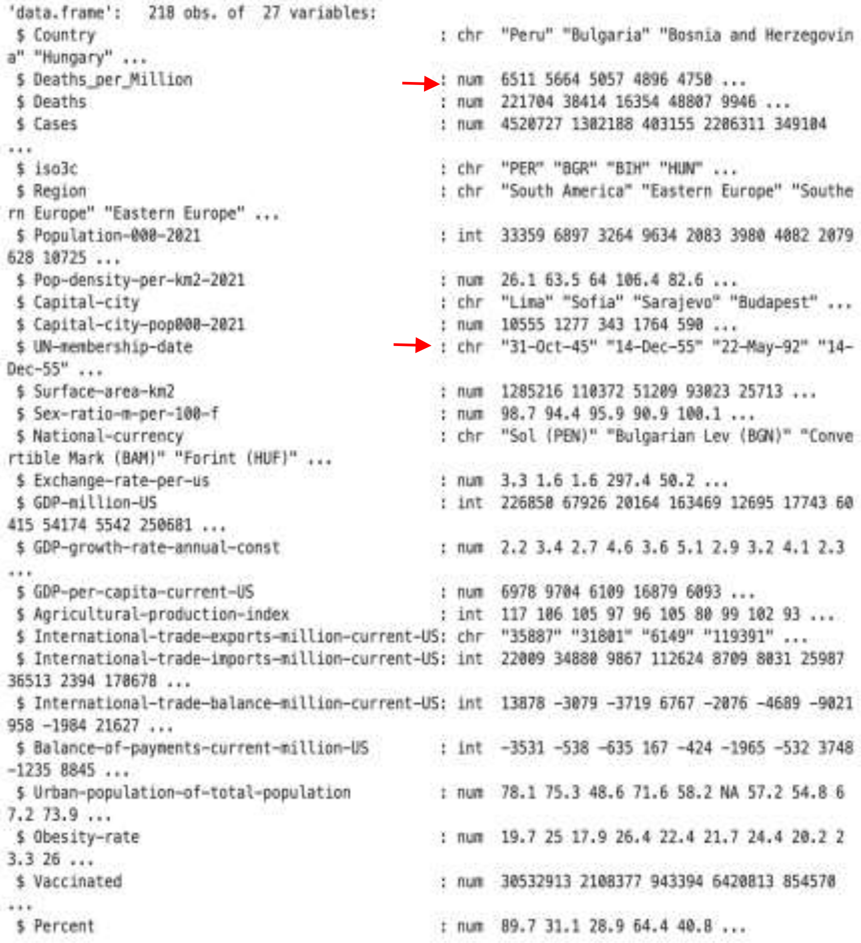
03 Columns are renamed to more descriptive names, making the dataframe more readable and easier to work with.

Data Transformation



```
 $ X                                                 : int  4 5 6 7 8 9 10 11 12 13 ...
 $ Country.x                                         : chr  "Peru" "Bulgaria" "Bosnia and Herzeg
ovina" "Hungary" ...
 $ Deaths...million                                  : chr  "6,511" "5,664" "5,057" "4,896" ...
 $ Deaths                                            : chr  "221,704" "38,414" "16,354" "48,807"
...
 $ Cases                                             : chr  "4,520,727" "1,302,188" "403,155"
"2,206,311" ...
 $ iso3c                                             : chr  "PER" "BGR" "BIH" "HUN" ...
 $ Country.y                                         : chr  "Peru" "Bulgaria" "Bosnia and Herzeg
ovina" "Hungary" ...
 $ Region.x                                          : chr  "South America" "Eastern Europe" "So
uthern Europe" "Eastern Europe" ...
 $ Population..000..2021.                            : int  33359 6897 3264 9634 2083 3980 4082
2079 628 10725 ...
 $ Pop..density..per.km2..2021.                      : num  26.1 63.5 64 106.4 82.6 ...
 $ Capital.city                                      : chr  "Lima" "Sofia" "Sarajevo" "Budapest"
...
 $ Capital.city.pop...000..2021.                     : num  10555 1277 343 1764 590 ...
 $ UN.membership.date                                : chr  "31-Oct-45" "14-Dec-55" "22-May-92"
"14-Dec-55" ...
 $ Surface.area..km2.                                : chr  "1285216" "110372" "51209" "93023"
...
 $ Sex.ratio..m.per.100.f.                           : num  98.7 94.4 95.9 90.9 100.1 ...
 $ National.currency                                 : chr  "Sol (PEN)" "Bulgarian Lev (BGN)" "C
onvertible Mark (BAM)" "Forint (HUF)" ...
 $ Exchange.rate..per.US..                           : num  3.3 1.6 1.6 297.4 50.2 ...
 $ GDP..Gross.domestic.product..million.current.US.. : int  226850 67926 20164 163469 12695 1774
3 60415 54174 5542 250681 ...
 $ GDP.growth.rate..annual....const..2015.prices.    : num  2.2 3.4 2.7 4.6 3.6 5.1 2.9 3.2 4.1
2.3 ...
 $ GDP.per.capita..current.US..                      : num  6978 9704 6109 16879 6093 ...
 $ Agricultural.production.index..2014.2016.100.     : int  117 106 105 97 96 105 80 99 102 93
...
 $ International.trade..exports..million.current.US..: chr  "35887" "31801" "6149" "119391" ...
 $ International.trade..imports..million.current.US..: int  22009 34880 9867 112624 8709 8031 25
987 36513 2394 170678 ...
 $ International.trade..balance..million.current.US..: int  13878 -3079 -3719 6767 -2076 -4689 -
9021 958 -1984 21627 ...
 $ Balance.of.payments..current.account..million.US..: int  -3531 -538 -635 167 -424 -1965 -532
3748 -1235 8845 ...
 $ Urban.population....of.total.population.          : num  78.1 75.3 48.6 71.6 58.2 NA 57.2 54.
8 67.2 73.9 ...
 $ Country                                           : chr  "Peru" "Bulgaria" "Bosnia and Herzeg
ovina" "Hungary" ...
 $ Obesity.rate....                                  : num  19.7 25 17.9 26.4 22.4 21.7 24.4 20.
2 23.3 26 ...
 $ Location                                          : chr  "Peru" "Bulgaria" "Bosnia and Herzeg
ovina" "Hungary" ...
 $ Vaccinated                                        : chr  "30,532,913" "2,108,377" "943,394"
"6,420,813" ...
 $ Percent                                           : num  89.7 31.1 28.9 64.4 40.8 ...
```

Year ambiguity



```
'data.frame':	218 obs. of  27 variables:
 $ Country                                      : chr  "Peru" "Bulgaria" "Bosnia and Herzegovin
a" "Hungary" ...
 $ Deaths_per_Million                           : num  6511 5664 5057 4896 4750 ...
 $ Deaths                                       : num  221704 38414 16354 48807 9946 ...
 $ Cases                                        : num  4520727 1302188 403155 2206311 349104
...
 $ iso3c                                        : chr  "PER" "BGR" "BIH" "HUN" ...
 $ Region                                       : chr  "South America" "Eastern Europe" "Southe
rn Europe" "Eastern Europe" ...
 $ Population-000-2021                          : int  33359 6897 3264 9634 2083 3980 4082 2079
628 10725 ...
 $ Pop-density-per-km2-2021                     : num  26.1 63.5 64 106.4 82.6 ...
 $ Capital-city                                 : chr  "Lima" "Sofia" "Sarajevo" "Budapest" ...
 $ Capital-city-pop000-2021                     : num  10555 1277 343 1764 590 ...
 $ UN-membership-date                           : chr  "31-Oct-45" "14-Dec-55" "22-May-92" "14-
Dec-55" ...
 $ Surface-area-km2                             : num  1285216 110372 51209 93023 25713 ...
 $ Sex-ratio-m-per-100-f                        : num  98.7 94.4 95.9 90.9 100.1 ...
 $ National-currency                            : chr  "Sol (PEN)" "Bulgarian Lev (BGN)" "Conve
rtible Mark (BAM)" "Forint (HUF)" ...
 $ Exchange-rate-per-us                         : num  3.3 1.6 1.6 297.4 50.2 ...
 $ GDP-million-US                               : int  226850 67926 20164 163469 12695 17743 60
415 54174 5542 250681 ...
 $ GDP-growth-rate-annual-const                 : num  2.2 3.4 2.7 4.6 3.6 5.1 2.9 3.2 4.1 2.3
...
 $ GDP-per-capita-current-US                    : num  6978 9704 6109 16879 6093 ...
 $ Agricultural-production-index                : int  117 106 105 97 96 105 80 99 102 93 ...
 $ International-trade-exports-million-current-US: chr  "35887" "31801" "6149" "119391" ...
 $ International-trade-imports-million-current-US: int  22009 34880 9867 112624 8709 8031 25987
36513 2394 170678 ...
 $ International-trade-balance-million-current-US: int  13878 -3079 -3719 6767 -2076 -4689 -9021
958 -1984 21627 ...
 $ Balance-of-payments-current-million-US       : int  -3531 -538 -635 167 -424 -1965 -532 3748
-1235 8845 ...
 $ Urban-population-of-total-population         : num  78.1 75.3 48.6 71.6 58.2 NA 57.2 54.8 6
7.2 73.9 ...
 $ Obesity-rate                                 : num  19.7 25 17.9 26.4 22.4 21.7 24.4 20.2 2
3.3 26 ...
 $ Vaccinated                                   : num  30532913 2108377 943394 6420813 854578
...
 $ Percent                                      : num  89.7 31.1 28.9 64.4 40.8 ...
```

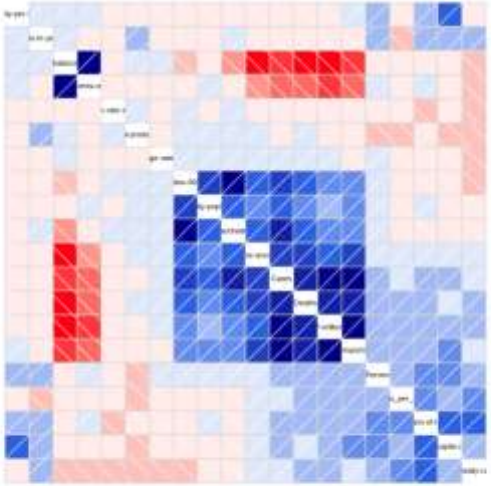
Analysis

Highly correlated variables:

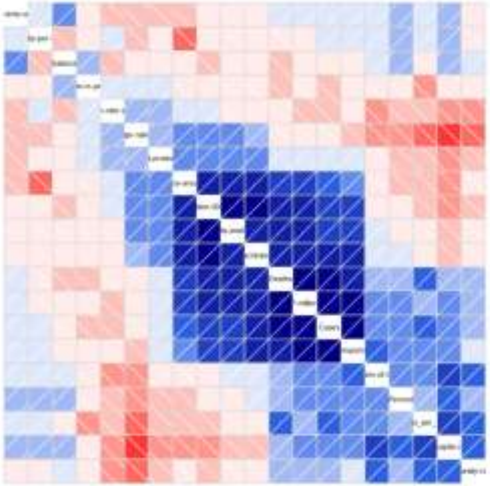
- Population-000-2021
- Capital-city-pop-000-2021
- Vaccinated
- Surface-area-km2
- Cases
- Deaths
- GDP-million-US
- International-trade-imports-million-current-US

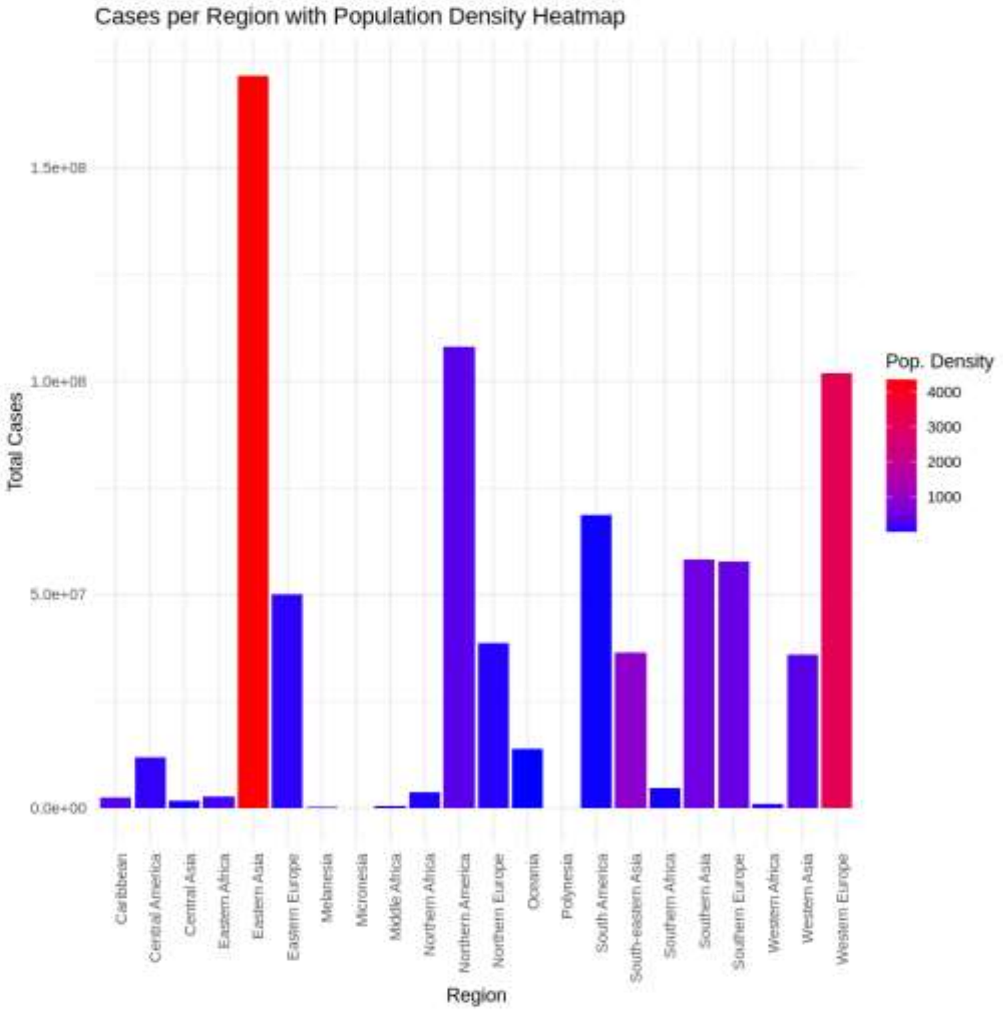
These variables have low missingess, which data would be more available and reliable for analysis.

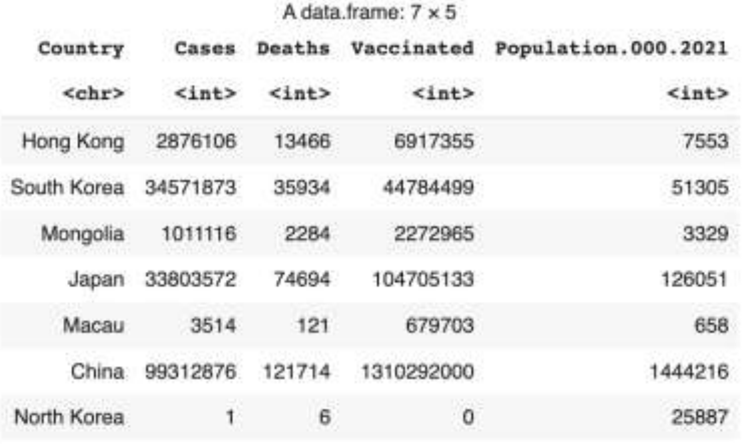
Correlation method: Pearson



Correlation method: Spearman

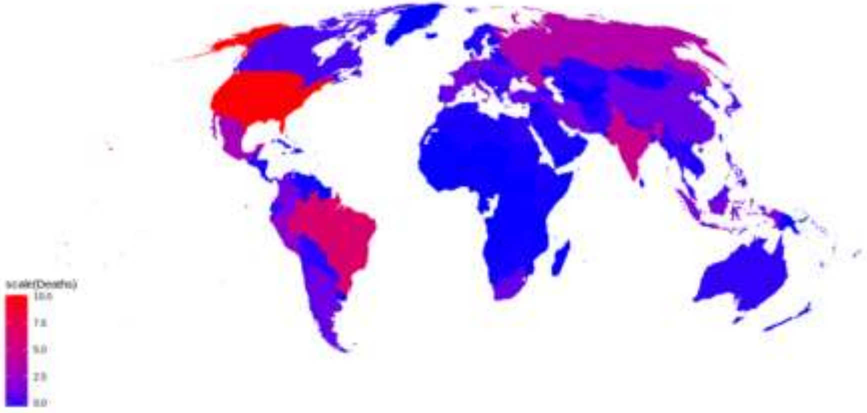


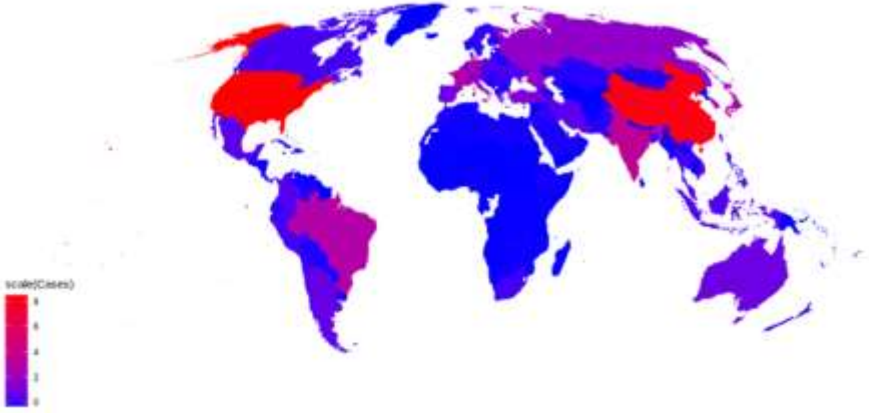


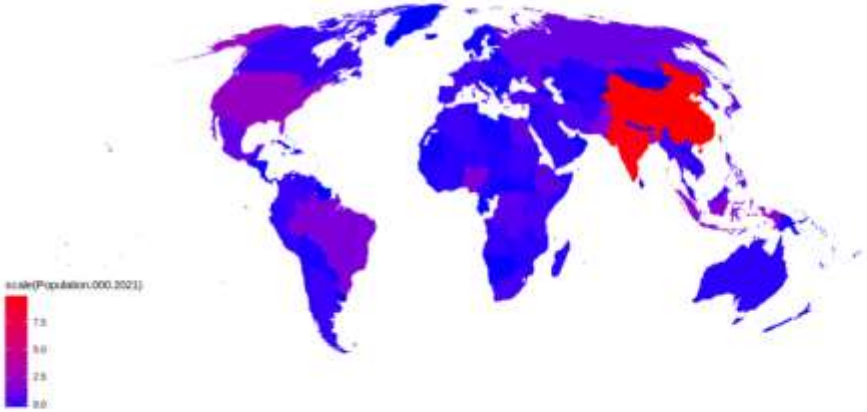


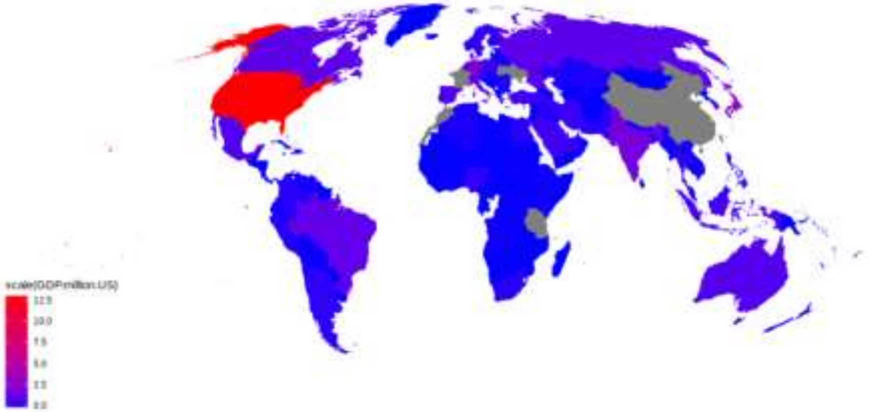
A data.frame: 7 × 5

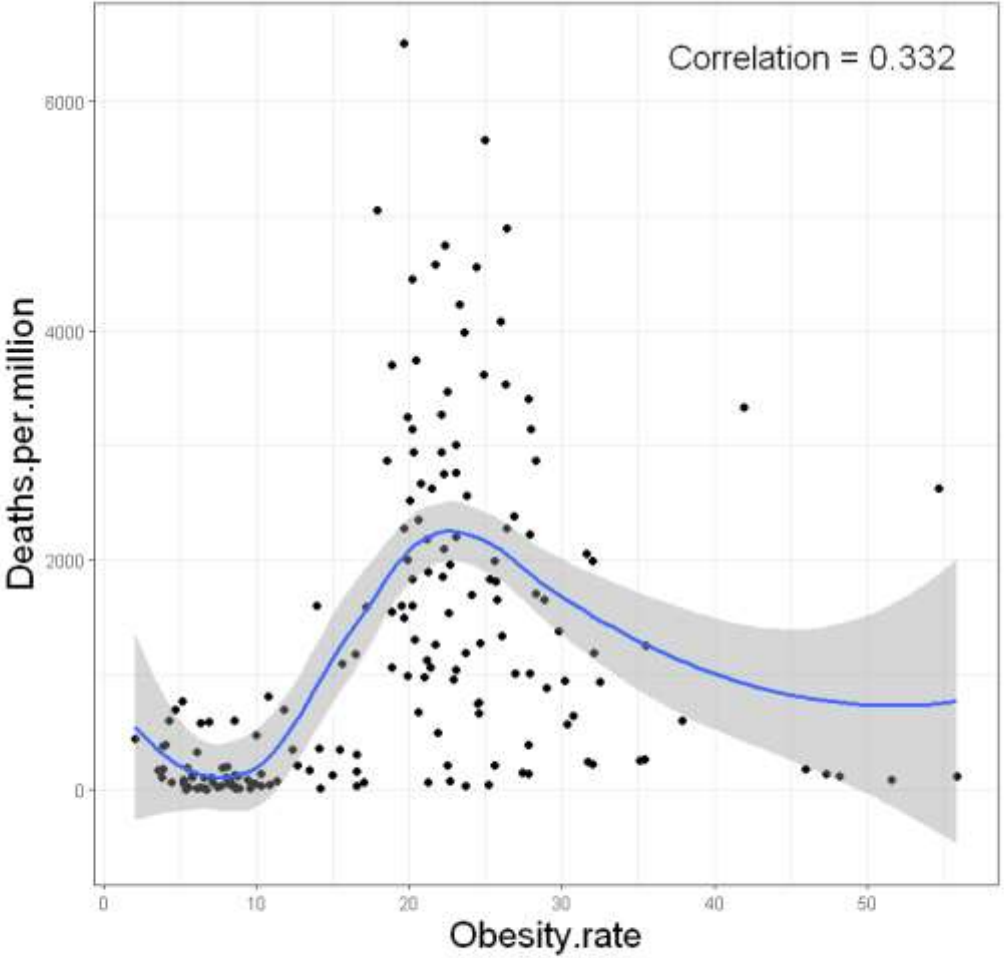
Country <chr>	Cases <int>	Deaths <int>	Vaccinated <int>	Population.000.2021 <int>
Hong Kong	2876106	13466	6917355	7553
South Korea	34571873	35934	44784499	51305
Mongolia	1011116	2284	2272965	3329
Japan	33803572	74694	104705133	126051
Macau	3514	121	679703	658
China	99312876	121714	1310292000	1444216
North Korea	1	6	0	25887


scale(Deaths)
10.0
7.5
5.0
2.5
0.0

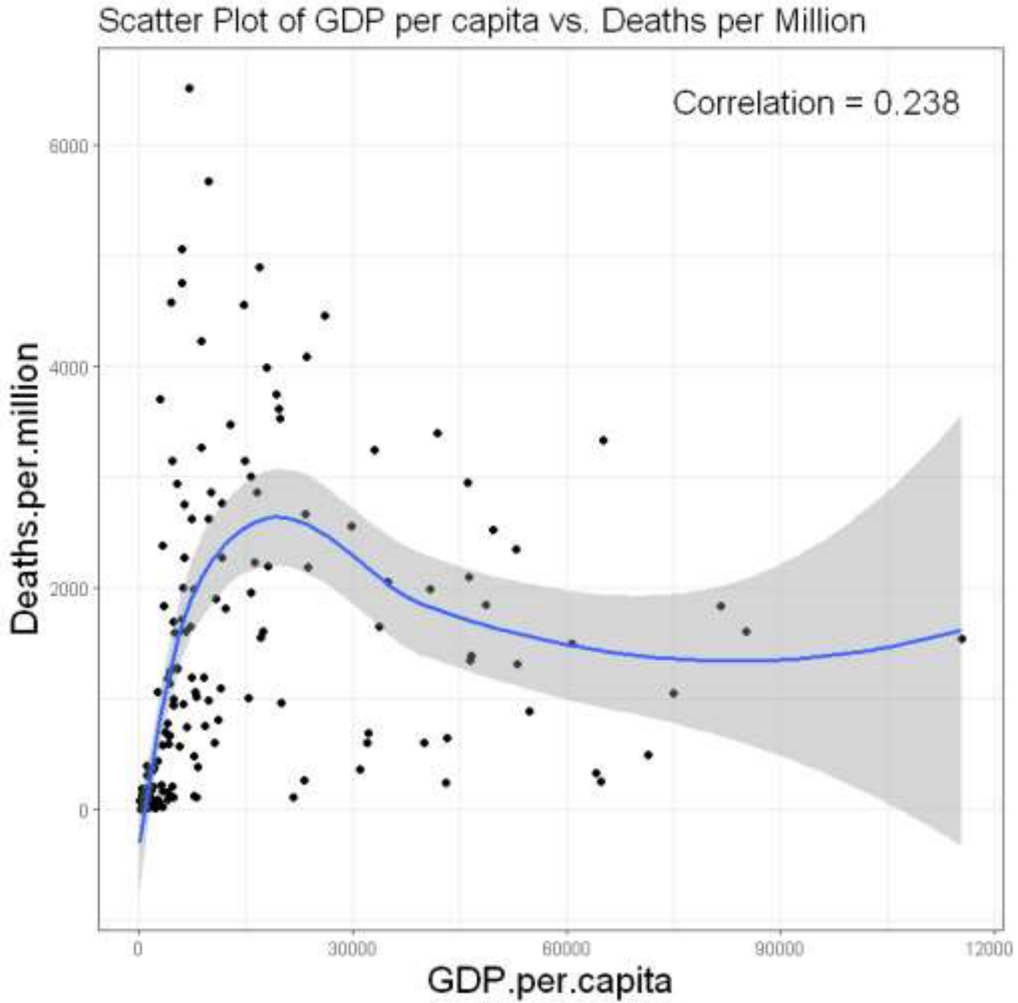

scale(Cases)
8
6
4
2
0


scale(Population.000.2021)
7.5
5.0
2.5
0.0


scale(GDPmillion.US)
12.5
10.0
7.5
5.0
2.5
0.0



Scatter Plot of Obesity Rate vs. Deaths per Million



Scatter Plot of GDP per capita vs. Deaths per Million