B518 | Week 4 | Group Project | Submission 1

Group 4 - Alex Toon | Nicholas Carlson | Divya Reddy Konda

2025-09-30

1 Project Idea One - Covid 19 (2021 ONLY - USA, UK, China, Belgium)

1.1 1) Original Source: URL: https://docs.owid.io/projects/covid/en/latest/dataset.html

1.2 2) Dataset moved: https://catalog.ourworldindata.org/garden/covid/latest/compact/compact.csv

1.2.1 Selection Criteria:

1.2.1.1 The data (see above) does meet the criteria of the assignment. In that it is relevant to health, publically accessible, sizable (61 columns and 530,292 rows), includes both categorical (e.g. country) and continuous variables (e.g. new_cases_per_million, total_deaths_per_million) and finally has been ethically sourced and de-identified.

```
## [1] "https://docs.owid.io/projects/covid/en/latest/dataset.html"
```

```
## n_rows n_cols
## 530292     61
```

```
##  [1] "country"                       "date"
##  [3] "total_cases"                   "new_cases"
##  [5] "new_cases_smoothed"            "total_cases_per_million"
##  [7] "new_cases_per_million"         "new_cases_smoothed_per_million"
##  [9] "total_deaths"                  "new_deaths"
## [11] "new_deaths_smoothed"           "total_deaths_per_million"
```

```
## [1] "Ukraine"                      "United Arab Emirates"
## [3] "United Kingdom"               "United States"
## [5] "United States Virgin Islands"
```

1.3 2) Introduction (4-6 Sentences)

This project uses Covid 19 data from 'Our world in data'. We use this to primarily compare how daily new cases per million varied across four countries in 2021. We focus on 2021 to keep our comparisons on a common phase of the pandemic. The dataset itself does cover many more countries and years and also includes data on total cases and total deaths. We used the fields that have the suffix 'per_million'as any comparisons scale by population size.

1.4 3) Why this dataset? (1 paragraph)

This dataset is highly relevant to health outcomes. The dataset is also very well documented, large (61 columns and 530k rows). For analysis potential, this dataset has both continuous fields (total_deaths_per_million, new_cases_per_million) and categorical (e.g. Country), feasible for tables, histograms, boxplots, time trends. Using the fields with the suffix "per_million" allows better scaling for cross country comparisons and summaries.

1.5 4) Variables and structure

```
## n_rows n_cols
## 530292     61
```

```
## [1] "country"                "date"
## [3] "total_cases"            "new_cases"
## [5] "new_cases_smoothed"     "total_cases_per_million"
```

```
##
## character   integer   logical   numeric
##         4        10         1        46
```

```
## 'data.frame':    530292 obs. of  61 variables:
##  $ country                                : chr  "Afghanistan" "Afghanistan" "Afghanistar
##  $ date                                   : chr  "2020-01-01" "2020-01-02" "2020-01-03" "
##  $ total_cases                            : int  NA NA NA 0 0 0 0 0 0 0 ...
##  $ new_cases                              : int  NA NA NA 0 0 0 0 0 0 0 ...
##  $ new_cases_smoothed                     : num  NA NA NA NA NA NA NA NA 0 0 ...
##  $ total_cases_per_million                : num  NA NA NA 0 0 0 0 0 0 0 ...
##  $ new_cases_per_million                  : num  NA NA NA 0 0 0 0 0 0 0 ...
##  $ new_cases_smoothed_per_million         : num  NA NA NA NA NA NA NA NA 0 0 ...
##  $ total_deaths                           : int  NA NA NA 0 0 0 0 0 0 0 ...
##  $ new_deaths                             : int  NA NA NA 0 0 0 0 0 0 0 ...
##  $ new_deaths_smoothed                    : num  NA NA NA NA NA NA NA NA 0 0 ...
##  $ total_deaths_per_million               : num  NA NA NA 0 0 0 0 0 0 0 ...
##  $ new_deaths_per_million                 : num  NA NA NA 0 0 0 0 0 0 0 ...
##  $ new_deaths_smoothed_per_million        : num  NA NA NA NA NA NA NA NA 0 0 ...
##  $ excess_mortality                       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ excess_mortality_cumulative            : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
##  $ excess_mortality_cumulative_absolute      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ excess_mortality_cumulative_per_million   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ hosp_patients                             : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ hosp_patients_per_million                 : num  NA NA NA NA NA NA NA NA NA NA ...
##   [list output truncated]
```

```
##        country       date total_cases new_cases new_cases_smoothed
## 1  Afghanistan 2020-01-01          NA        NA                 NA
## 2  Afghanistan 2020-01-02          NA        NA                 NA
## 3  Afghanistan 2020-01-03          NA        NA                 NA
## 4  Afghanistan 2020-01-04           0         0                 NA
## 5  Afghanistan 2020-01-05           0         0                 NA
## 6  Afghanistan 2020-01-06           0         0                 NA
## 7  Afghanistan 2020-01-07           0         0                 NA
## 8  Afghanistan 2020-01-08           0         0                 NA
## 9  Afghanistan 2020-01-09           0         0                  0
## 10 Afghanistan 2020-01-10           0         0                  0
##    total_cases_per_million new_cases_per_million new_cases_smoothed_per_million
## 1                       NA                    NA                              NA
## 2                       NA                    NA                              NA
## 3                       NA                    NA                              NA
## 4                        0                     0                              NA
## 5                        0                     0                              NA
## 6                        0                     0                              NA
## 7                        0                     0                              NA
## 8                        0                     0                              NA
## 9                        0                     0                               0
## 10                       0                     0                               0
##    total_deaths new_deaths new_deaths_smoothed total_deaths_per_million
## 1            NA         NA                  NA                       NA
## 2            NA         NA                  NA                       NA
## 3            NA         NA                  NA                       NA
## 4             0          0                  NA                        0
## 5             0          0                  NA                        0
## 6             0          0                  NA                        0
## 7             0          0                  NA                        0
## 8             0          0                  NA                        0
## 9             0          0                   0                        0
## 10            0          0                   0                        0
##    new_deaths_per_million new_deaths_smoothed_per_million excess_mortality
## 1                      NA                              NA              NA
## 2                      NA                              NA              NA
## 3                      NA                              NA              NA
## 4                       0                              NA              NA
## 5                       0                              NA              NA
## 6                       0                              NA              NA
## 7                       0                              NA              NA
## 8                       0                              NA              NA
```

```
## 9                       0                                0              NA
## 10                      0                                0              NA
##    excess_mortality_cumulative excess_mortality_cumulative_absolute
## 1                           NA                                    NA
## 2                           NA                                    NA
## 3                           NA                                    NA
## 4                           NA                                    NA
## 5                           NA                                    NA
## 6                           NA                                    NA
## 7                           NA                                    NA
## 8                           NA                                    NA
## 9                           NA                                    NA
## 10                          NA                                    NA
##    excess_mortality_cumulative_per_million hosp_patients
## 1                                       NA            NA
## 2                                       NA            NA
## 3                                       NA            NA
## 4                                       NA            NA
## 5                                       NA            NA
## 6                                       NA            NA
## 7                                       NA            NA
## 8                                       NA            NA
## 9                                       NA            NA
## 10                                      NA            NA
##    hosp_patients_per_million weekly_hosp_admissions
## 1                         NA                    NA
## 2                         NA                    NA
## 3                         NA                    NA
## 4                         NA                    NA
## 5                         NA                    NA
## 6                         NA                    NA
## 7                         NA                    NA
## 8                         NA                    NA
## 9                         NA                    NA
## 10                        NA                    NA
##    weekly_hosp_admissions_per_million icu_patients icu_patients_per_million
## 1                                  NA           NA                       NA
## 2                                  NA           NA                       NA
## 3                                  NA           NA                       NA
## 4                                  NA           NA                       NA
## 5                                  NA           NA                       NA
## 6                                  NA           NA                       NA
## 7                                  NA           NA                       NA
## 8                                  NA           NA                       NA
## 9                                  NA           NA                       NA
## 10                                 NA           NA                       NA
##    weekly_icu_admissions weekly_icu_admissions_per_million stringency_index
## 1                     NA                                NA                0
```

```
## 2                    NA                                  NA                0
## 3                    NA                                  NA                0
## 4                    NA                                  NA                0
## 5                    NA                                  NA                0
## 6                    NA                                  NA                0
## 7                    NA                                  NA                0
## 8                    NA                                  NA                0
## 9                    NA                                  NA                0
## 10                   NA                                  NA                0
##    reproduction_rate total_tests new_tests total_tests_per_thousand
## 1                 NA          NA        NA                       NA
## 2                 NA          NA        NA                       NA
## 3                 NA          NA        NA                       NA
## 4                 NA          NA        NA                       NA
## 5                 NA          NA        NA                       NA
## 6                 NA          NA        NA                       NA
## 7                 NA          NA        NA                       NA
## 8                 NA          NA        NA                       NA
## 9                 NA          NA        NA                       NA
## 10                NA          NA        NA                       NA
##    new_tests_per_thousand new_tests_smoothed new_tests_smoothed_per_thousand
## 1                      NA                 NA                               NA
## 2                      NA                 NA                               NA
## 3                      NA                 NA                               NA
## 4                      NA                 NA                               NA
## 5                      NA                 NA                               NA
## 6                      NA                 NA                               NA
## 7                      NA                 NA                               NA
## 8                      NA                 NA                               NA
## 9                      NA                 NA                               NA
## 10                     NA                 NA                               NA
##    positive_rate tests_per_case total_vaccinations people_vaccinated
## 1             NA             NA                 NA                NA
## 2             NA             NA                 NA                NA
## 3             NA             NA                 NA                NA
## 4             NA             NA                 NA                NA
## 5             NA             NA                 NA                NA
## 6             NA             NA                 NA                NA
## 7             NA             NA                 NA                NA
## 8             NA             NA                 NA                NA
## 9             NA             NA                 NA                NA
## 10            NA             NA                 NA                NA
##    people_fully_vaccinated total_boosters new_vaccinations
## 1                       NA             NA                NA
## 2                       NA             NA                NA
## 3                       NA             NA                NA
## 4                       NA             NA                NA
## 5                       NA             NA                NA
```

```
## 6                       NA             NA                NA
## 7                       NA             NA                NA
## 8                       NA             NA                NA
## 9                       NA             NA                NA
## 10                      NA             NA                NA
##    new_vaccinations_smoothed total_vaccinations_per_hundred
## 1                         NA                              NA
## 2                         NA                              NA
## 3                         NA                              NA
## 4                         NA                              NA
## 5                         NA                              NA
## 6                         NA                              NA
## 7                         NA                              NA
## 8                         NA                              NA
## 9                         NA                              NA
## 10                        NA                              NA
##    people_vaccinated_per_hundred people_fully_vaccinated_per_hundred
## 1                             NA                                   NA
## 2                             NA                                   NA
## 3                             NA                                   NA
## 4                             NA                                   NA
## 5                             NA                                   NA
## 6                             NA                                   NA
## 7                             NA                                   NA
## 8                             NA                                   NA
## 9                             NA                                   NA
## 10                            NA                                   NA
##    total_boosters_per_hundred new_vaccinations_smoothed_per_million
## 1                          NA                                   NA
## 2                          NA                                   NA
## 3                          NA                                   NA
## 4                          NA                                   NA
## 5                          NA                                   NA
## 6                          NA                                   NA
## 7                          NA                                   NA
## 8                          NA                                   NA
## 9                          NA                                   NA
## 10                         NA                                   NA
##    new_people_vaccinated_smoothed new_people_vaccinated_smoothed_per_hundred
## 1                               NA                                         NA
## 2                               NA                                         NA
## 3                               NA                                         NA
## 4                               NA                                         NA
## 5                               NA                                         NA
## 6                               NA                                         NA
## 7                               NA                                         NA
## 8                               NA                                         NA
## 9                               NA                                         NA
```

```
## 10                            NA                                      NA
##    code continent population population_density median_age life_expectancy
## 1   AFG      Asia   40578847           62.21555     16.752          65.617
## 2   AFG      Asia   40578847           62.21555     16.752          65.617
## 3   AFG      Asia   40578847           62.21555     16.752          65.617
## 4   AFG      Asia   40578847           62.21555     16.752          65.617
## 5   AFG      Asia   40578847           62.21555     16.752          65.617
## 6   AFG      Asia   40578847           62.21555     16.752          65.617
## 7   AFG      Asia   40578847           62.21555     16.752          65.617
## 8   AFG      Asia   40578847           62.21555     16.752          65.617
## 9   AFG      Asia   40578847           62.21555     16.752          65.617
## 10  AFG      Asia   40578847           62.21555     16.752          65.617
##    gdp_per_capita extreme_poverty diabetes_prevalence handwashing_facilities
## 1         1516.273              NA                10.9               48.21469
## 2         1516.273              NA                10.9               48.21469
## 3         1516.273              NA                10.9               48.21469
## 4         1516.273              NA                10.9               48.21469
## 5         1516.273              NA                10.9               48.21469
## 6         1516.273              NA                10.9               48.21469
## 7         1516.273              NA                10.9               48.21469
## 8         1516.273              NA                10.9               48.21469
## 9         1516.273              NA                10.9               48.21469
## 10        1516.273              NA                10.9               48.21469
##    hospital_beds_per_thousand human_development_index
## 1                        0.39                      NA
## 2                        0.39                      NA
## 3                        0.39                      NA
## 4                        0.39                      NA
## 5                        0.39                      NA
## 6                        0.39                      NA
## 7                        0.39                      NA
## 8                        0.39                      NA
## 9                        0.39                      NA
## 10                       0.39                      NA

##    country              date             total_cases          new_cases
##  Length:530292      Length:530292      Min.   :        0   Min.   :      0
##  Class :character   Class :character   1st Qu.:     9106   1st Qu.:      0
##  Mode  :character   Mode  :character   Median :    84204   Median :      0
##                                        Mean   : 14120121   Mean   :  10590
##                                        3rd Qu.:  1067030   3rd Qu.:     64
##                                        Max.   :778523540   Max.   :8401906
##                                        NA's   :13866       NA's   :17148
##  new_cases_smoothed  total_cases_per_million new_cases_per_million
##  Min.   :      0.0   Min.   :     0          Min.   :0.000e+00
##  1st Qu.:      0.0   1st Qu.:  2988          1st Qu.:0.000e+00
##  Median :      4.7   Median : 46164          Median :0.000e+00
##  Mean   :  10615.0   Mean   :129928          Mean   :9.928e+01
```

```
##  3rd Qu.:    198.4   3rd Qu.:192015          3rd Qu.:7.427e+00
##  Max.   :6402033.0   Max.   :769807          Max.   :2.308e+05
##  NA's   :18353       NA's   :13866           NA's   :17148
##  new_cases_smoothed_per_million  total_deaths       new_deaths
##  Min.   :    0.000               Min.   :      0   Min.   :    0.0
##  1st Qu.:    0.000               1st Qu.:     68   1st Qu.:    0.0
##  Median :    0.872               Median :   1020   Median :    0.0
##  Mean   :   99.514               Mean   : 154260   Mean   :  101.2
##  3rd Qu.:   32.357               3rd Qu.:  12283   3rd Qu.:    0.0
##  Max.   :37463.746               Max.   :7100783   Max.   :57167.0
##  NA's   :18353                   NA's   :13866     NA's   :16277
##  new_deaths_smoothed total_deaths_per_million new_deaths_per_million
##  Min.   :    0.000   Min.   :   0.00          Min.   :  0.0000
##  1st Qu.:    0.000   1st Qu.:  40.66          1st Qu.:  0.0000
##  Median :    0.000   Median : 390.36          Median :  0.0000
##  Mean   :  101.435   Mean   : 921.58          Mean   :  0.6188
##  3rd Qu.:    1.857   3rd Qu.:1440.10          3rd Qu.:  0.0000
##  Max.   :14820.714   Max.   :6603.65          Max.   :608.6427
##  NA's   :17489       NA's   :13866            NA's   :16277
##  new_deaths_smoothed_per_million excess_mortality excess_mortality_cumulative
##  Min.   :  0.0000                Min.   :-95.92   Min.   :-44.23
##  1st Qu.:  0.0000                1st Qu.: -1.51   1st Qu.:  2.19
##  Median :  0.0000                Median :  5.59   Median :  8.23
##  Mean   :  0.6203                Mean   : 10.87   Mean   :  9.79
##  3rd Qu.:  0.1951                3rd Qu.: 15.50   3rd Qu.: 15.15
##  Max.   :129.0729                Max.   :378.55   Max.   : 78.08
##  NA's   :17489                   NA's   :516619   NA's   :516619
##  excess_mortality_cumulative_absolute excess_mortality_cumulative_per_million
##  Min.   : -37726.1                    Min.   :-2936.4
##  1st Qu.:    185.8                    1st Qu.:  130.8
##  Median :   6558.0                    Median : 1313.0
##  Mean   :  55285.1                    Mean   : 1785.7
##  3rd Qu.:  38780.6                    3rd Qu.: 2873.2
##  Max.   :1349776.2                    Max.   :10293.5
##  NA's   :516619                       NA's   :516653
##  hosp_patients    hosp_patients_per_million weekly_hosp_admissions
##  Min.   :     0   Min.   :   0.00           Min.   :     0
##  1st Qu.:   186   1st Qu.:  31.00           1st Qu.:   223
##  Median :   776   Median :  74.24           Median :   864
##  Mean   :  3912   Mean   : 125.99           Mean   :  4292
##  3rd Qu.:  3051   3rd Qu.: 159.76           3rd Qu.:  3893
##  Max.   :154497   Max.   :1526.85           Max.   :153977
##  NA's   :489636   NA's   :489636            NA's   :505795
##  weekly_hosp_admissions_per_million  icu_patients    icu_patients_per_million
##  Min.   :  0.00                     Min.   :    0    Min.   :  0.00
##  1st Qu.: 23.73                     1st Qu.:   21    1st Qu.:  2.33
##  Median : 56.28                     Median :   90    Median :  6.43
##  Mean   : 82.62                     Mean   :  661    Mean   : 15.66
```

```
##  3rd Qu.:110.00                    3rd Qu.:  413    3rd Qu.: 18.78
##  Max.   :717.08                    Max.   :28891    Max.   :180.68
##  NA's   :505795                    NA's   :491176   NA's   :491176
##  weekly_icu_admissions weekly_icu_admissions_per_million stringency_index
##  Min.   :   0.0        Min.   :  0.00                    Min.   :  0.00
##  1st Qu.:  17.0        1st Qu.:  1.55                    1st Qu.: 22.22
##  Median :  92.0        Median :  4.64                    Median : 42.59
##  Mean   : 317.9        Mean   :  9.67                    Mean   : 42.68
##  3rd Qu.: 353.0        3rd Qu.: 12.65                    3rd Qu.: 62.04
##  Max.   :4838.0        Max.   :224.98                    Max.   :100.00
##  NA's   :519299        NA's   :519299                    NA's   :327532
##  reproduction_rate  total_tests          new_tests
##  Min.   :-0.07     Min.   :0.000e+00   Min.   :       1
##  1st Qu.: 0.71     1st Qu.:3.647e+05   1st Qu.:    2244
##  Median : 0.95     Median :2.067e+06   Median :    8783
##  Mean   : 0.91     Mean   :2.110e+07   Mean   :   67285
##  3rd Qu.: 1.14     3rd Qu.:1.025e+07   3rd Qu.:   37229
##  Max.   : 5.87     Max.   :9.214e+09   Max.   :35855632
##  NA's   :344609    NA's   :450905      NA's   :454889
##  total_tests_per_thousand new_tests_per_thousand new_tests_smoothed
##  Min.   :    0.00         Min.   :  0.00         Min.   :       0
##  1st Qu.:   43.59         1st Qu.:  0.29         1st Qu.:    1486
##  Median :  234.14         Median :  0.97         Median :    6570
##  Mean   :  924.25         Mean   :  3.27         Mean   :  142178
##  3rd Qu.:  894.37         3rd Qu.:  2.91         3rd Qu.:   32205
##  Max.   :32925.82         Max.   :531.06         Max.   :14769984
##  NA's   :450905           NA's   :454889         NA's   :426327
##  new_tests_smoothed_per_thousand positive_rate    tests_per_case
##  Min.   :  0.00                  Min.   : 0.00    Min.   :     0.00
##  1st Qu.:  0.20                  1st Qu.: 1.61    1st Qu.:     6.83
##  Median :  0.85                  Median : 5.49    Median :    17.60
##  Mean   :  2.83                  Mean   :10.42    Mean   :  1833.14
##  3rd Qu.:  2.58                  3rd Qu.:14.27    3rd Qu.:    58.51
##  Max.   :147.60                  Max.   :94.51    Max.   :789603.80
##  NA's   :426327                  NA's   :428732   NA's   :429647
##  total_vaccinations  people_vaccinated   people_fully_vaccinated
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:1.893e+06   1st Qu.:1.011e+06   1st Qu.:8.626e+05
##  Median :1.574e+07   Median :6.980e+06   Median :6.801e+06
##  Mean   :5.837e+08   Mean   :2.641e+08   Mean   :2.430e+08
##  3rd Qu.:1.292e+08   3rd Qu.:5.497e+07   3rd Qu.:5.269e+07
##  Max.   :1.372e+10   Max.   :5.645e+09   Max.   :5.198e+09
##  NA's   :447070      NA's   :451281      NA's   :453281
##  total_boosters      new_vaccinations   new_vaccinations_smoothed
##  Min.   :0.000e+00   Min.   :       0   Min.   :       0
##  1st Qu.:3.853e+05   1st Qu.:    1784   1st Qu.:     216
##  Median :6.124e+06   Median :   22056   Median :    3250
##  Mean   :1.431e+08   Mean   :  766640   Mean   :  275096
```

```
##  3rd Qu.:4.441e+07   3rd Qu.:  196541   3rd Qu.:   29708
##  Max.   :2.841e+09   Max.   :49673200   Max.   :43691812
##  NA's   :471508      NA's   :461416     NA's   :327452
##  total_vaccinations_per_hundred people_vaccinated_per_hundred
##  Min.   :  0.00                 Min.   :  0.00
##  1st Qu.: 42.99                 1st Qu.: 26.79
##  Median :126.98                 Median : 63.22
##  Mean   :121.55                 Mean   : 52.67
##  3rd Qu.:193.10                 3rd Qu.: 76.75
##  Max.   :415.88                 Max.   :112.08
##  NA's   :447070                 NA's   :451281
##  people_fully_vaccinated_per_hundred total_boosters_per_hundred
##  Min.   :  0.00                      Min.   :  0.00
##  1st Qu.: 19.48                      1st Qu.:  2.03
##  Median : 56.59                      Median : 27.94
##  Mean   : 47.26                      Mean   : 31.14
##  3rd Qu.: 73.06                      3rd Qu.: 53.80
##  Max.   :110.19                      Max.   :140.53
##  NA's   :453281                      NA's   :471508
##  new_vaccinations_smoothed_per_million new_people_vaccinated_smoothed
##  Min.   :     0.00                     Min.   :0.000e+00
##  1st Qu.:    97.04                     1st Qu.:1.940e+01
##  Median :   513.02                     Median :5.660e+02
##  Mean   :  1728.33                     Mean   :1.001e+05
##  3rd Qu.:  2160.33                     3rd Qu.:8.047e+03
##  Max.   :117342.85                     Max.   :2.107e+07
##  NA's   :327452                        NA's   :327452
##  new_people_vaccinated_smoothed_per_hundred     code
##  Min.   : 0.00                               Length:530292
##  1st Qu.: 0.00                               Class :character
##  Median : 0.01                               Mode  :character
##  Mean   : 0.07
##  3rd Qu.: 0.06
##  Max.   :11.73
##  NA's   :327452
##   continent           population        population_density    median_age
##  Length:530292      Min.   :5.130e+02   Min.   :1.365e-01   Min.   :14.30
##  Class :character   1st Qu.:4.554e+05   1st Qu.:3.627e+01   1st Qu.:22.24
##  Mode  :character   Median :6.035e+06   Median :9.208e+01   Median :31.68
##                     Mean   :1.315e+08   Mean   :3.805e+02   Mean   :31.18
##                     3rd Qu.:2.972e+07   3rd Qu.:2.375e+02   3rd Qu.:39.08
##                     Max.   :8.021e+09   Max.   :2.134e+04   Max.   :59.88
##                     NA's   :17098       NA's   :25401       NA's   :23320
##  life_expectancy gdp_per_capita     extreme_poverty  diabetes_prevalence
##  Min.   :18.82   Min.   :   708.2   Min.   : 0.000   Min.   : 1.100
##  1st Qu.:68.75   1st Qu.:  5155.6   1st Qu.: 0.498   1st Qu.: 5.600
##  Median :74.70   Median : 14740.0   Median : 2.817   Median : 7.400
##  Mean   :73.45   Mean   : 22520.6   Mean   :14.425   Mean   : 9.058
```

```
##  3rd Qu.:78.79   3rd Qu.: 34663.5   3rd Qu.:22.002   3rd Qu.:11.100
##  Max.   :85.75   Max.   :117747.0   Max.   :87.740   Max.   :30.800
##  NA's   :21246   NA's   :116142     NA's   :190395   NA's   :82901
##  handwashing_facilities hospital_beds_per_thousand human_development_index
##  Min.   :  3.44         Min.   : 0.300             Mode:logical
##  1st Qu.: 26.20         1st Qu.: 1.320             NA's:530292
##  Median : 70.15         Median : 2.600
##  Mean   : 59.25         Mean   : 3.183
##  3rd Qu.: 88.47         3rd Qu.: 4.260
##  Max.   :100.00         Max.   :13.800
##  NA's   :290110         NA's   :209535
```

1.6 5) Research questions

- 1. What share of days exceed a threshold (to simulated a government policy threshold to "flatten the curve") e.g 50 cases per million in each country
- 2. Which of the selected countries had the highest typical daily new cases per million in 2021
- 3. How did the monthly mean of new cases per million over 2021 for each country

1.7 6) Data clean up & Processing plan

We parsed the date field and derived a 'year' variable, then restricted the dataset to 2021 to keep figures more legible and comparable. We fixed our analysis to a small set of countries (United States, United Kingdom, China, Belgium) and then verified each has sufficient non missing values for 'new_cases_per_million' in 2021. this processing prepares the data for descriptive statistics and many visualisations.

```
##  United States United Kingdom          China        Belgium
##            365             365            365            365
```

1.8 7) Descriptive statistics & visualisations

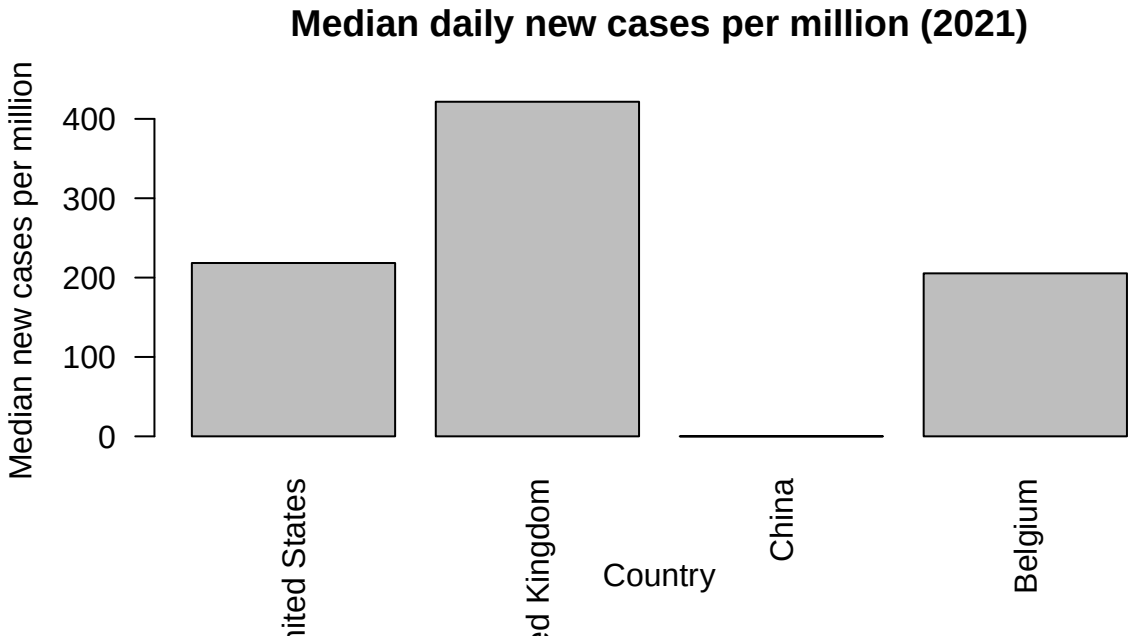
We summarise categories (counts/proportions), report center & spread for one mumeric variable and add simple plots to visualise patterns ### 7.1) One-Way frequency table (categorical) Counts and proportions for a categorical variable

```
##
##  United States United Kingdom          China        Belgium
##            365             365            365            365
```

```
##
##  United States United Kingdom          China        Belgium
##           0.25            0.25           0.25           0.25
```

1.8.1 7.2) Bar Chart of Disease.Category (counts)

Bar chart / Bar plot of disease category by count



1.8.2 7.3) Two way table (category by category)

```
##
##                  FALSE TRUE
##   United States     37  328
##   United Kingdom    55  310
##   China            365    0
##   Belgium           20  345
```

```
##
##                  FALSE  TRUE
##   United States  0.101 0.899
##   United Kingdom 0.151 0.849
##   China          1.000 0.000
##   Belgium        0.055 0.945
```

```
##
##                  FALSE  TRUE
##   United States  0.078 0.334
##   United Kingdom 0.115 0.315
##   China          0.765 0.000
##   Belgium        0.042 0.351
```

```
##
##                   FALSE TRUE  Sum
##   United States      37  328  365
##   United Kingdom     55  310  365
##   China             365    0  365
##   Belgium            20  345  365
##   Sum               477  983 1460
```

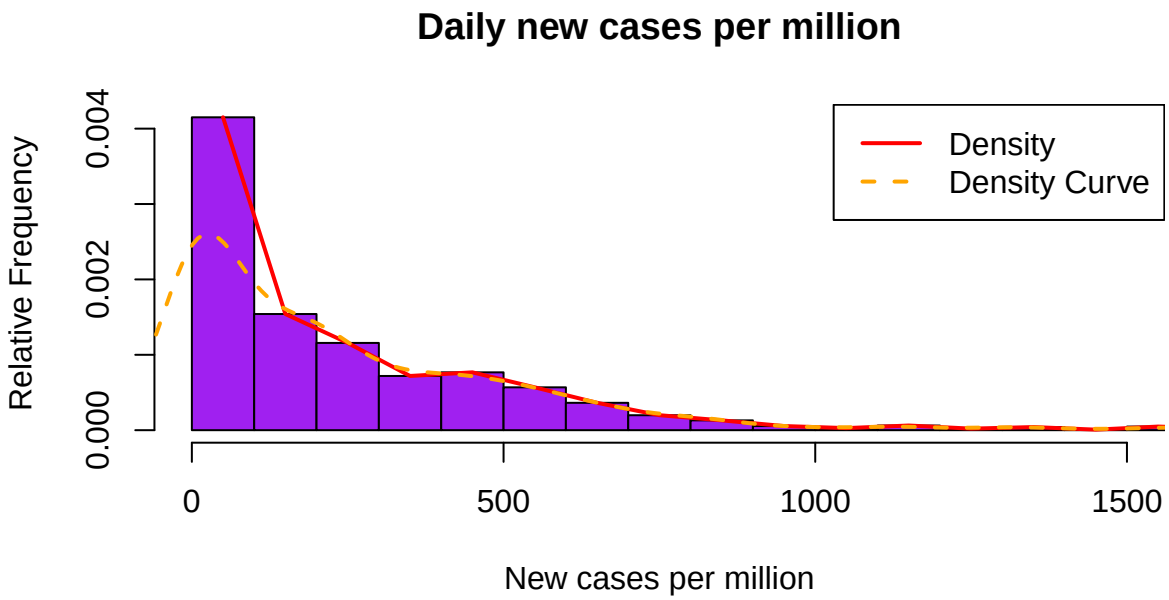
```
##  United States United Kingdom          China        Belgium
##          0.899          0.849          0.000          0.945
```

1.8.3 7.4) Center & Spread (overall, selected countries, 2021)

```
## median    IQR     sd
##  162.9  379.7  357.4
```

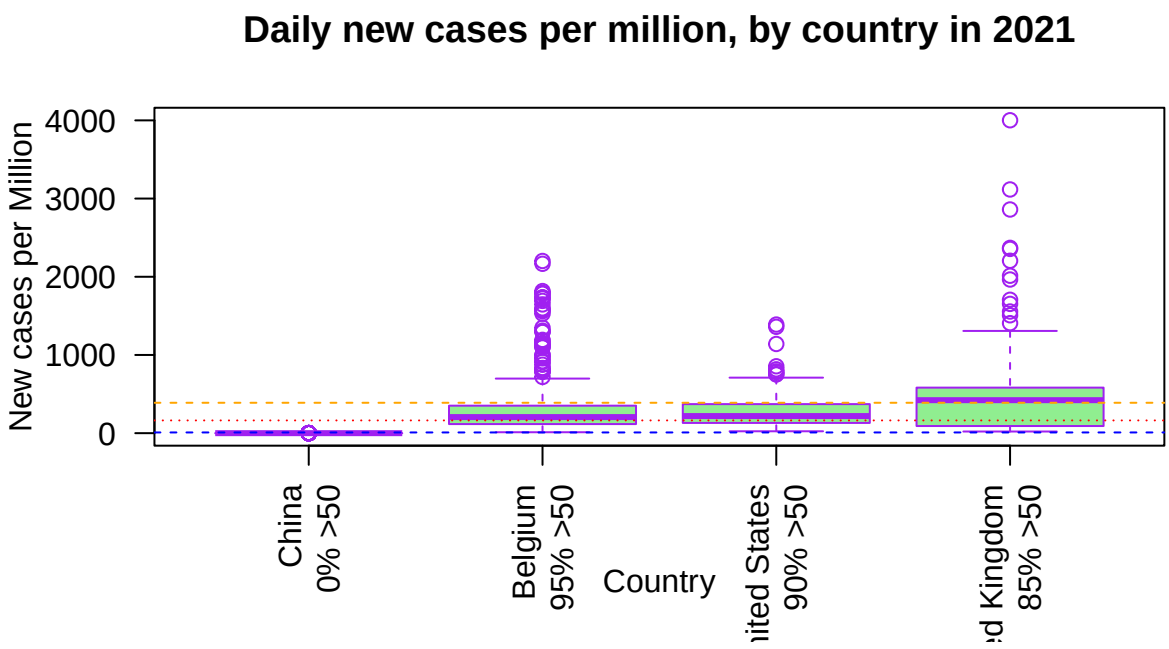
```
##          country median   IQR    sd
## 1  United States  218.4 240.7 201.6
## 2 United Kingdom  421.6 491.1 456.4
## 3          China    0.0   0.1   0.1
## 4        Belgium  205.4 235.4 396.2
```

1.8.4 7.5) Histogram (shape of the distrubution)

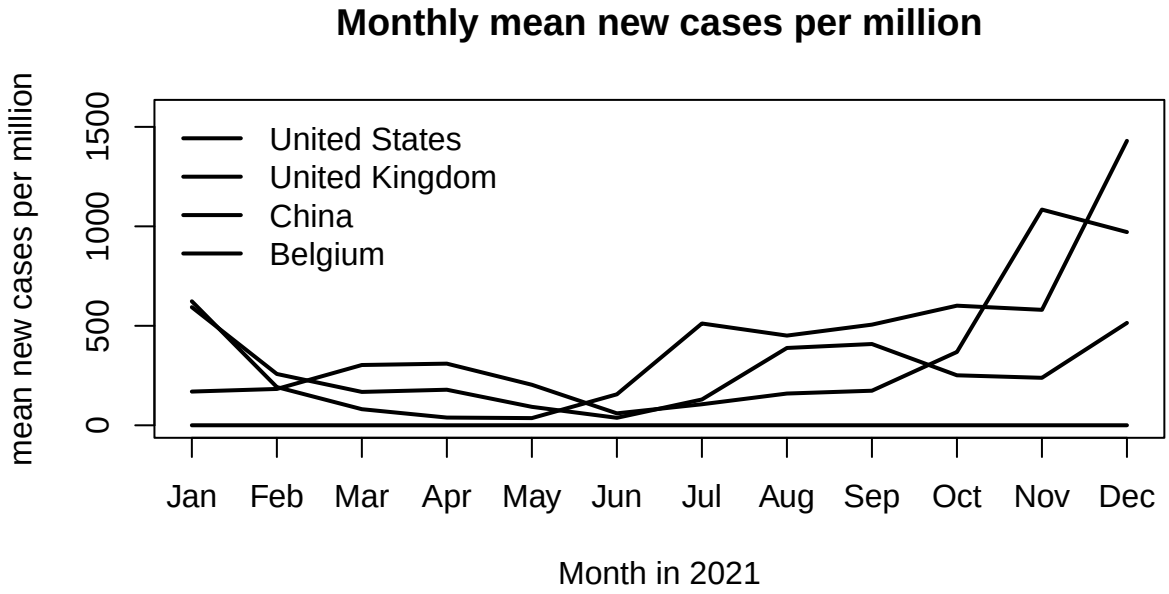


1.8.5 7.6) Boxplot (numeric by category)

Boxplot (Mortality rate by category)



1.8.6 7.7) Simple time trend (average by year)



1.9 8) Planned statisical methods

We will compare 2021 distrubutions of 'new_cases_per_million' across the four countries using….

1.10 9) Limitations

- Measurement differences - countries have different reporting rules, testing cadence & breadth.
- Scope - Only 2021 was analysed. Other years or waves of the disease may show other patterns.
- per million rates do not adjust for demographics of each country, which may show other patterns.
- China has several near zero analysis - This may reflect reporting practices of this specific country

2 Project Idea Two - Covid 19 Hospitalizations in France

3 Link to the dataset

Kaggle - Coronavirusdataset France (file: `chiffres-cles.csv`)
Actual URL: https://www.kaggle.com/datasets/mclikmb4/coronavirusdataset-france?select=chiffres-cles.csv Google drive URL: https://drive.google.com/file/d/1rXHdGEDWFAMaitmkNSgehAt_e2FaC_PZ/view?usp=sharing

4 Introduction to the dataset

This dataset provides daily COVID-19 surveillance indicators for France at multiple geographic granularities (country, region, department, overseas collectivities). Each record includes a calendar date, a location code, and a location name, enabling comparisons across space and time. Indicators cover hospitalized patients, ICU occupancy, cumulative deaths, cumulative recoveries, and daily flows of new admissions (hospital and ICU). Source/provenance fields support auditability. The structure suits descriptive analyses and visualizations, with optional regional comparisons to highlight spatial heterogeneity. These indicators and their definitions are documented on the Kaggle dataset page (mclikmb4, 2020-2021).

5 Dataset justification

Relevance: Directly biomedical/public-health, reflecting real-world hospital and ICU loads during COVID-19.

Size/structure: The file far exceeds the minimum requirements (well over 100 rows and more than 20 columns) and includes both categorical (granularity, location IDs, sources) and continuous (counts) variables.

Accessibility/ethics: Publicly accessible aggregated, de-identified counts suitable for academic use.

Analytical potential: Enables trend estimation, wave identification, geographic comparison, and lead-lag analysis between admissions ("flow") and occupancy ("stock").

Ethical use. The dataset consists of aggregated, de-identified counts without PII; no patient-level identifiers are present, aligning with course requirements for ethical, public data.

6 Variables description

Key columns:
`date` (daily), `granularity` (country, region, department), `location_code` (location code), `location_name` (location name).

Indicators:
- `hospitalized` - current hospitalized patients
- `icu_patients` - current ICU patients
- `deaths` - cumulative deaths
- `recovered` - cumulative recoveries
- `new_hospitalizations` - new daily hospital admissions
- `new_icu_admissions` - new daily ICU admissions

Additional fields:
`confirmed_cases` and `tested` may be present with different levels of completeness.
Note: Due to several missing/invalid values (NaN/Inf), the `tested` column is largely unusable for analysis and is excluded from primary summaries and plots.

Source metadata:
`source_name`, `source_url`, `source_archive`, `source_type`.

Table 1: Row counts by geographic granularity

granularity	n
department	40715
region	7708
country	817
overseas_collectivity	131
world	83

Table 2: Summary statistics for key numeric indicators

variable	n	mean	sd	median	min	max
confirmed_cases	3081	121010.685	508142.429	27.0	0	3560764
deaths	47928	920.086	4150.452	135.0	0	70574
hospitalized	46826	578.225	2597.057	91.0	0	33497
icu_patients	46743	80.489	387.667	10.0	0	7148
new_hospitalizations	46095	32.664	166.648	4.0	0	4281
new_icu_admissions	46095	5.421	28.033	0.0	0	771
recovered	46712	3949.800	17835.138	645.5	0	299624

variable	n	mean	sd	median	min	max
tested	0	NaN	NA	NA	Inf	-Inf

7 Research question(s)

1. **National waves:** How did France's national hospitalization and ICU occupancy evolve across early pandemic waves (2020-2021)?

2. **Flow-stock timing:** Do peaks in new hospital admissions precede peaks in current hospitalizations, and by roughly how many days?

8 Data cleanup and processing plan

- **Parsing and types:** Ensure the `date` field is properly parsed as a date variable and convert indicator fields into numeric types for consistency.
- **Subsetting:** For national trends, include only rows classified as country with `location_code` = "FRA". For geographic comparisons, restrict the dataset to rows where `granularity` is region.
- **Missingness:** Quantify missing values for each column and handle them transparently by applying listwise deletion for plotted series (no imputation).
- **Duplicates:** Identify and remove duplicate entries defined by the combination of `date` and `location_code`.
- **Provenance:** Retain all source metadata fields, and include them in the appendix when relevant for transparency.

9 Descriptive statistics (figures in Appendix)

France's national indicators exhibit multi-wave patterns during 2020-2021. Hospital occupancy and ICU burden rise and fall in tandem with case surges, while cumulative deaths increase monotonically. The timing relationship between new admissions (flow) and current occupancy (stock) suggests admissions lead occupancy by several days. For visuals supporting these statements, see Appendix Figures A1-A3. Tables above summarize structure and central tendencies.

Across all rows, the median current hospitalizations was 91, with an IQR of 25-285; ICU occupancy had a much lower median, which is expected since ICU is a subset of the total hospital (median 10), consistent with ICU being a subset of total hospital burden.

10 Planned statistical methods

- **Lagged cross-correlation** between `new_hospitalizations` (flow) and `hospitalized` (stock) to estimate lead time from admissions to occupancy.
- **Regional comparison** of ICU vs hospital burden by wave period (medians, IQRs).
- **Simple time-series decomposition** on national hospitalizations to separate trend/seasonal/residual components (if applicable).

11 Limitations

Several fields like `tested` and early `confirmed_cases` have bad coverage over time, and indicators are hospital-centric rather than community-representative. Counts are aggregated and de-identified, so patient-level cannot be controlled. Because the dataset mixes granularities (national, regional, departmental), comparing across levels requires careful subsetting (`granularity == "country"` for national trends). These constraints limit causal interpretation, so we have to focus more on descriptive trends and clearly labeled comparisons.

12 Appendix

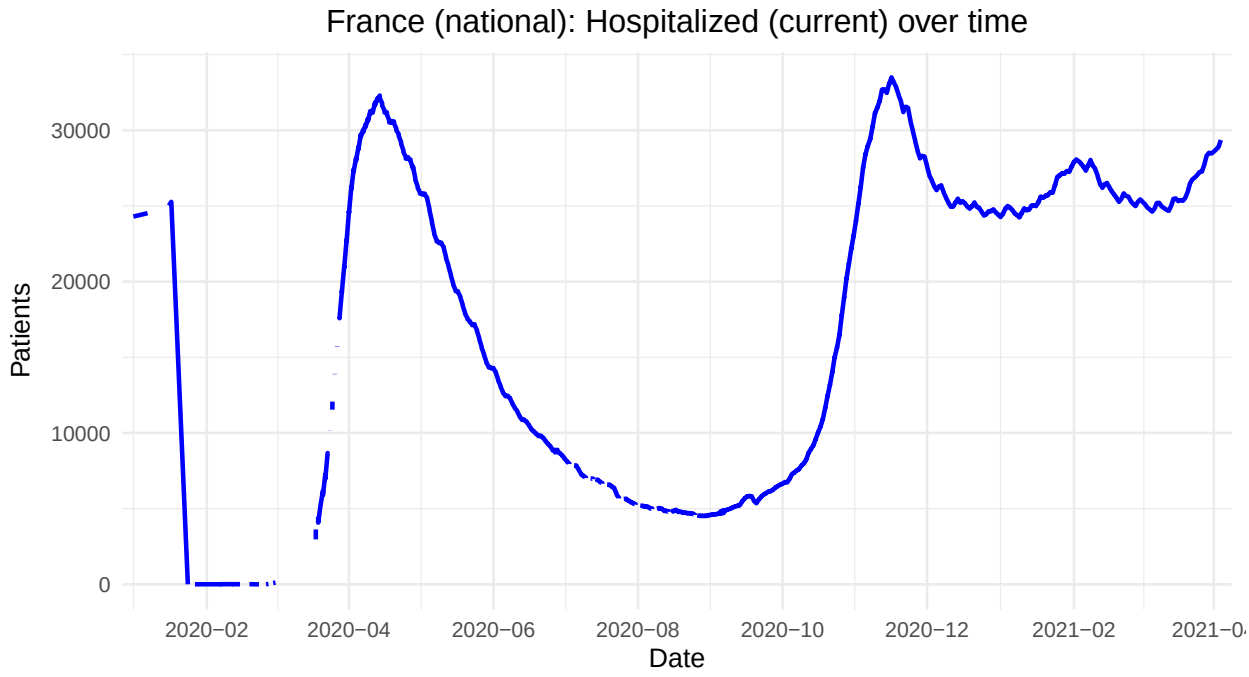


Figure 1: France (national): Hospitalized (current) over time.

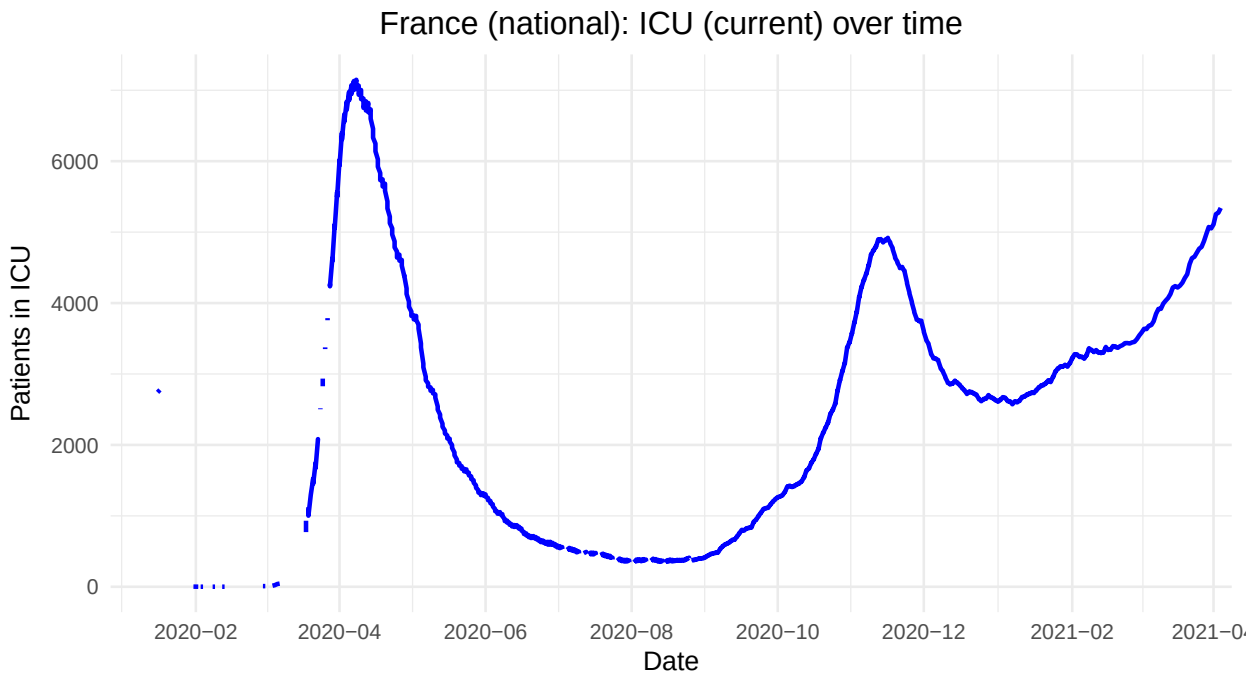


Figure 2: France (national): ICU (current) over time.

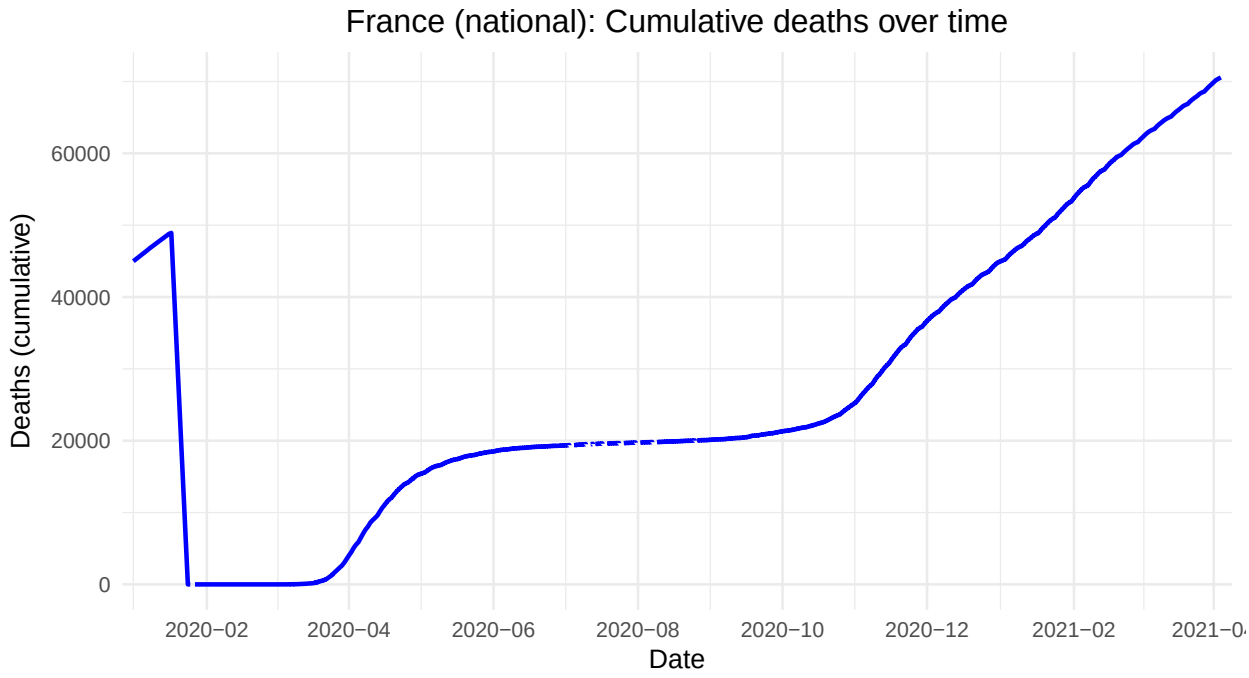


Figure 3: France (national): Cumulative deaths over time.

13 Project Idea Three - Heart attack

14 1.Link to the dataset

https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

15 2. Introduction to dataset

The Heart Attack Prediction Dataset, available on Kaggle, is a comprehensive resource for studying the clinical, lifestyle, and demographic factors associated with cardiovascular risk. It consists of 8,763 de-identified patient records, including continuous variables such as age, cholesterol, blood pressure, and heart rate, as well as categorical features like sex, chest pain type, smoking habits, diabetes status, and dietary patterns. Socioeconomic and geographic attributes, including income and region, further enrich the dataset by adding broader context to heart health predictors. The primary outcome variable indicates whether a patient is at risk of a heart attack, making the dataset well-suited for statistical analysis, visualization, and classification tasks. Its diverse mix of variables supports exploration of correlations, risk factors, and group comparisons, while also providing an ethical and accessible foundation for predictive modeling in cardiovascular health research.

16 3. Dataset justification

I chose the Heart Attack Prediction Dataset because it directly addresses a critical biomedical challenge cardiovascular disease which remains one of the leading causes of mortality worldwide. The dataset integrates clinical, lifestyle, and demographic variables, making it highly relevant for exploring the multifactorial nature of heart health. With its balanced mix of categorical and continuous features, it offers strong potential for applying a variety of statistical methods, visualizations, and predictive modeling techniques. Its size and diversity of attributes make it complex enough to yield meaningful insights, yet still manageable for academic analysis. Overall, this dataset provides both real-world relevance and analytical richness, making it an excellent candidate for this project.

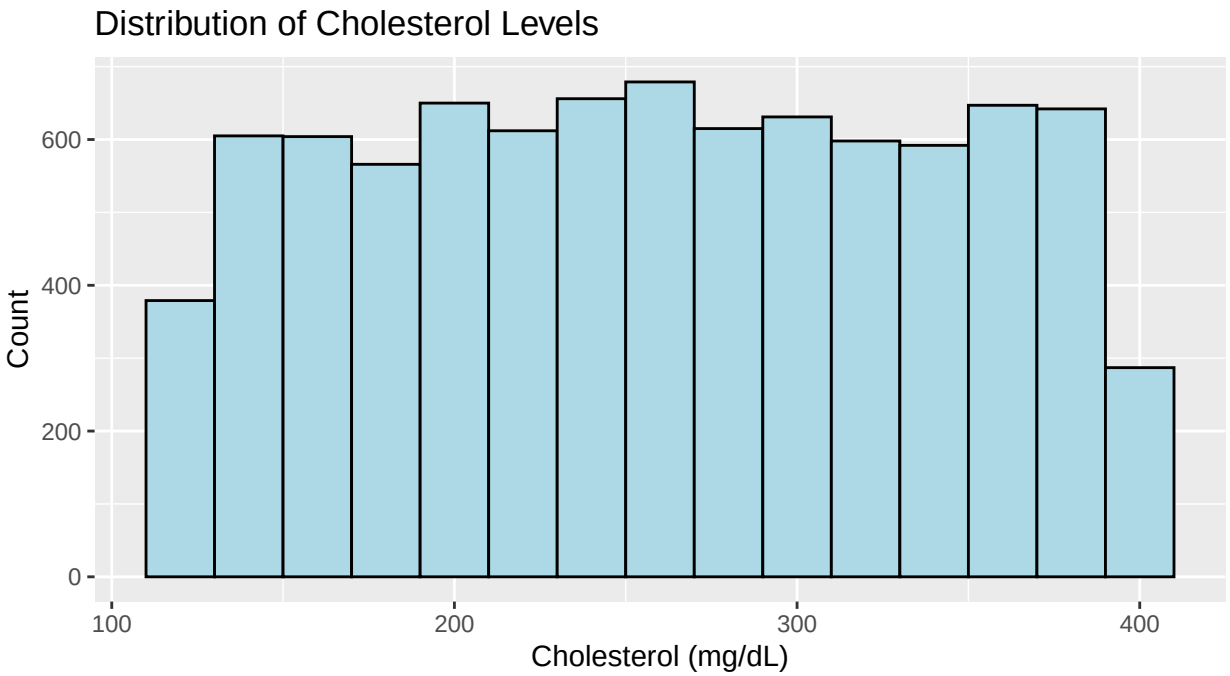
17 4. Variables description

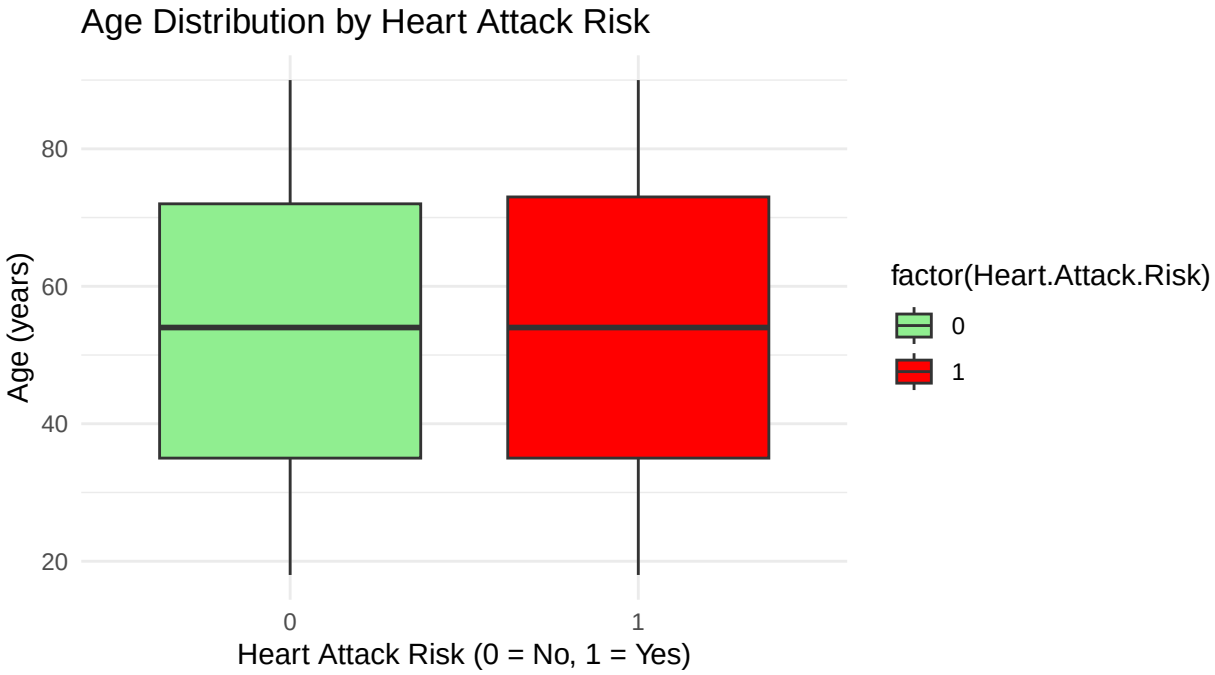
Key columns include Patient ID (unique identifier for each record), Age (in years), Sex (male or female), Cholesterol (cholesterol levels in mg/dL), Blood Pressure (systolic/diastolic in mmHg), Heart Rate (beats per minute), and BMI (body mass index, kg/m^2). Clinical indicators capture Diabetes status (Yes/No), Family History of heart problems (1 = Yes, 0 = No), Previous Heart Problems (1 = Yes, 0 = No), Medication Use (1 = Yes, 0 = No), and Triglyceride levels (mg/dL). Lifestyle-related attributes include Smoking (1 = Smoker, 0 = Non-smoker), Obesity (1 = Obese, 0 = Not obese), Alcohol Consumption (None, Light, Moderate, Heavy), Diet (Healthy, Average, Unhealthy), Exercise Hours Per Week, Physical Activity Days Per Week, Stress Level (1–10 scale), Sedentary Hours Per Day, and Sleep Hours Per Day. Socioeconomic and demographic fields consist of Income, Country, Continent, and Hemisphere. The target variable, Heart Attack Risk, is a binary indicator (1 = Yes, 0 = No) denoting whether the patient is at risk of a heart attack.

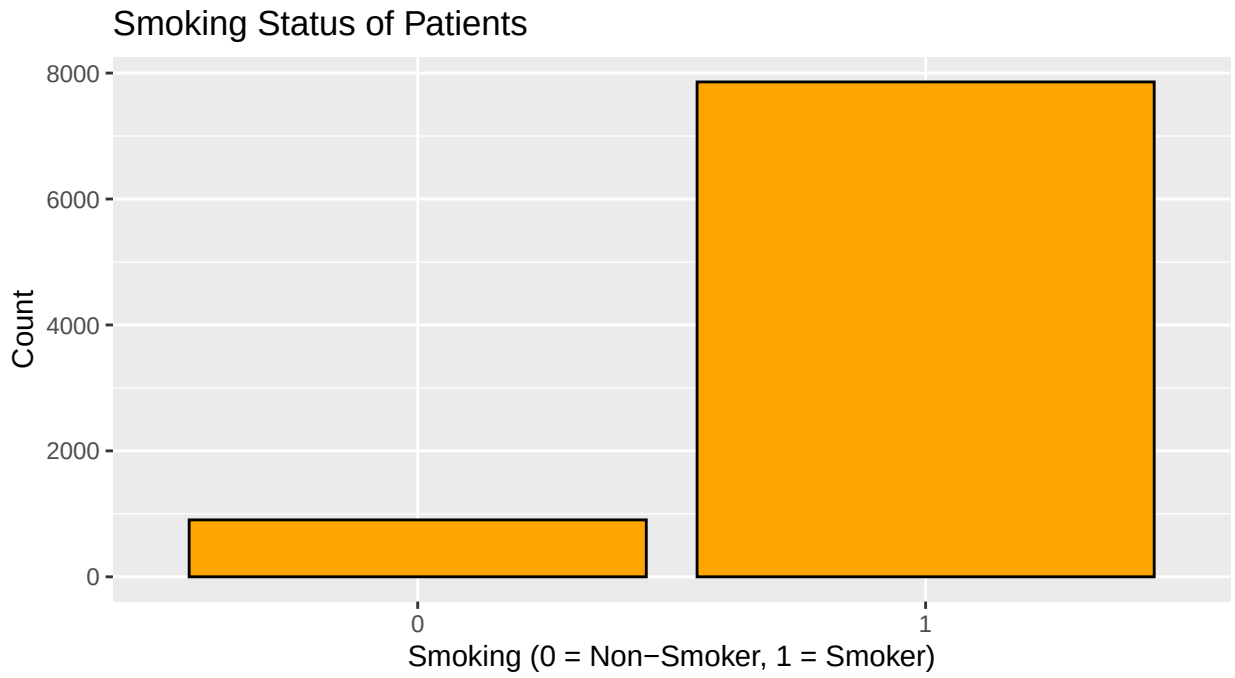
```
## 'data.frame':    8763 obs. of  26 variables:
##  $ Patient.ID                     : chr  "BMW7812" "CZE1114" "BNI9906" "JLN3497" ...
##  $ Age                            : int  67 21 21 84 66 54 90 84 20 43 ...
##  $ Sex                            : chr  "Male" "Male" "Female" "Male" ...
##  $ Cholesterol                    : int  208 389 324 383 318 297 358 220 145 248 ...
##  $ Blood.Pressure                 : chr  "158/88" "165/93" "174/99" "163/100" ...
##  $ Heart.Rate                     : int  72 98 72 73 93 48 84 107 68 55 ...
##  $ Diabetes                       : int  0 1 1 1 1 1 0 0 1 0 ...
##  $ Family.History                 : int  0 1 0 1 1 1 0 0 0 1 ...
##  $ Smoking                        : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ Obesity                        : int  0 1 0 0 1 0 0 1 1 1 ...
##  $ Alcohol.Consumption            : int  0 1 0 1 0 1 1 1 0 1 ...
##  $ Exercise.Hours.Per.Week        : num  4.17 1.81 2.08 9.83 5.8 ...
##  $ Diet                           : chr  "Average" "Unhealthy" "Healthy" "Average" ...
##  $ Previous.Heart.Problems        : int  0 1 1 1 1 1 0 0 0 0 ...
##  $ Medication.Use                 : int  0 0 1 0 0 1 0 1 0 0 ...
##  $ Stress.Level                   : int  9 1 9 9 6 2 7 4 5 4 ...
##  $ Sedentary.Hours.Per.Day        : num  6.62 4.96 9.46 7.65 1.51 ...
##  $ Income                         : int  261404 285768 235282 125640 160555 241339 190450 1
##  $ BMI                            : num  31.3 27.2 28.2 36.5 21.8 ...
##  $ Triglycerides                  : int  286 235 587 378 231 795 284 370 790 232 ...
##  $ Physical.Activity.Days.Per.Week: int  0 1 4 3 1 5 4 6 7 7 ...
##  $ Sleep.Hours.Per.Day            : int  6 7 4 4 5 10 10 7 4 7 ...
##  $ Country                        : chr  "Argentina" "Canada" "France" "Canada" ...
##  $ Continent                      : chr  "South America" "North America" "Europe" "North Am
##  $ Hemisphere                     : chr  "Southern Hemisphere" "Northern Hemisphere" "North
##  $ Heart.Attack.Risk              : int  0 0 0 0 0 1 1 1 0 0 ...

##   Patient.ID             Age            Sex             Cholesterol
##  Length:8763        Min.   :18.00   Length:8763        Min.   :120.0
##  Class :character   1st Qu.:35.00   Class :character   1st Qu.:192.0
##  Mode  :character   Median :54.00   Mode  :character   Median :259.0
##                     Mean   :53.71                      Mean   :259.9
##                     3rd Qu.:72.00                      3rd Qu.:330.0
##                     Max.   :90.00                      Max.   :400.0
##  Blood.Pressure       Heart.Rate        Diabetes      Family.History
##  Length:8763        Min.   : 40.00   Min.   :0.0000   Min.   :0.000
##  Class :character   1st Qu.: 57.00   1st Qu.:0.0000   1st Qu.:0.000
##  Mode  :character   Median : 75.00   Median :1.0000   Median :0.000
##                     Mean   : 75.02   Mean   :0.6523   Mean   :0.493
##                     3rd Qu.: 93.00   3rd Qu.:1.0000   3rd Qu.:1.000
##                     Max.   :110.00   Max.   :1.0000   Max.   :1.000
##     Smoking          Obesity       Alcohol.Consumption Exercise.Hours.Per.Week
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000      Min.   : 0.002442
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000      1st Qu.: 4.981579
##  Median :1.0000   Median :1.0000   Median :1.0000      Median :10.069559
##  Mean   :0.8968   Mean   :0.5014   Mean   :0.5981      Mean   :10.014284
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000      3rd Qu.:15.050018
```

```
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000      Max.   :19.998709
##      Diet           Previous.Heart.Problems Medication.Use    Stress.Level
##  Length:8763        Min.   :0.0000          Min.   :0.0000   Min.   : 1.00
##  Class :character   1st Qu.:0.0000          1st Qu.:0.0000   1st Qu.: 3.00
##  Mode  :character   Median :0.0000          Median :0.0000   Median : 5.00
##                     Mean   :0.4958          Mean   :0.4983   Mean   : 5.47
##                     3rd Qu.:1.0000          3rd Qu.:1.0000   3rd Qu.: 8.00
##                     Max.   :1.0000          Max.   :1.0000   Max.   :10.00
##  Sedentary.Hours.Per.Day     Income            BMI        Triglycerides
##  Min.   : 0.001263       Min.   : 20062   Min.   :18.00   Min.   : 30.0
##  1st Qu.: 2.998794       1st Qu.: 88310   1st Qu.:23.42   1st Qu.:225.5
##  Median : 5.933622       Median :157866   Median :28.77   Median :417.0
##  Mean   : 5.993690       Mean   :158263   Mean   :28.89   Mean   :417.7
##  3rd Qu.: 9.019125       3rd Qu.:227749   3rd Qu.:34.32   3rd Qu.:612.0
##  Max.   :11.999313       Max.   :299954   Max.   :40.00   Max.   :800.0
##  Physical.Activity.Days.Per.Week Sleep.Hours.Per.Day   Country
##  Min.   :0.00                    Min.   : 4.000      Length:8763
##  1st Qu.:2.00                    1st Qu.: 5.000      Class :character
##  Median :3.00                    Median : 7.000      Mode  :character
##  Mean   :3.49                    Mean   : 7.024
##  3rd Qu.:5.00                    3rd Qu.: 9.000
##  Max.   :7.00                    Max.   :10.000
##   Continent          Hemisphere        Heart.Attack.Risk
##  Length:8763        Length:8763        Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :0.0000
##                                        Mean   :0.3582
##                                        3rd Qu.:1.0000
##                                        Max.   :1.0000
```


Distribution of Cholesterol Levels
Count
600
400
200
0
100
200
300
400
Cholesterol (mg/dL)


Age Distribution by Heart Attack Risk
Age (years)
80
60
40
20
0
1
Heart Attack Risk (0 = No, 1 = Yes)
factor(Heart.Attack.Risk)
0
1



18 5. Research questions

1. Which clinical, lifestyle, and demographic factors are most strongly associated with the risk of heart attack in patients?
2. Which features contribute most to a machine learning model's decision boundary for predicting heart attack risk?

19 6. Data cleanup and processing plan

- Check for missing values: Identify NAs using colSums(is.na(hd)); if very few, remove those rows; if moderate, impute using mean/median for continuous variables (e.g., Cholesterol, BMI) and mode for categorical variables (e.g., Diet, Alcohol Consumption).
- Remove duplicate entries: Drop exact duplicates or repeated Patient IDs to avoid overrepresentation using hd <- hd[!duplicated(hd),].
- Fix inconsistent formats: Split Blood Pressure into two numeric columns (Systolic and Diastolic) and convert binary indicators (0/1) like Diabetes, Smoking, and Heart Attack Risk into categorical factors.
- Validate ranges & handle outliers: Review continuous variables (e.g., Cholesterol, Triglycerides, BMI, Sleep Hours) for biologically implausible values; correct, cap, or remove extreme outliers as appropriate.
- Standardize categorical variables: Ensure consistent levels for Sex (Male/Female), Diet (Healthy/Average/Unhealthy), and Alcohol Consumption (None/Light/Moderate/Heavy).

- Create derived variables: Add new groupings such as Age Groups (e.g., 18–30, 31–50, 51–70, 71–90) and BMI Categories (Underweight, Normal, Overweight, Obese) to facilitate group comparisons in descriptive statistics and visualization.

20 7. Descriptive statistics and data visualizations

```
##
## Variable: Age
## Mean: 53.70798
## Median: 54
## Range: 72
## Standard Deviation: 21.24951
##
## Variable: Cholesterol
## Mean: 259.8772
## Median: 259
## Range: 280
## Standard Deviation: 80.86328
##
## Variable: Heart.Rate
## Mean: 75.02168
## Median: 75
## Range: 70
## Standard Deviation: 20.55095
##
## Variable: BMI
## Mean: 28.89145
## Median: 28.769
## Range: 21.99487
## Standard Deviation: 6.319181
##
## Variable: Triglycerides
## Mean: 417.6771
## Median: 417
## Range: 770
## Standard Deviation: 223.7481
##
## Variable: Exercise.Hours.Per.Week
## Mean: 10.01428
## Median: 10.06956
## Range: 19.99627
## Standard Deviation: 5.783745
##
## Variable: Stress.Level
## Mean: 5.469702
## Median: 5
## Range: 9
```

```
## Standard Deviation: 2.859622
##
## Variable: Sedentary.Hours.Per.Day
## Mean: 5.99369
## Median: 5.933622
## Range: 11.99805
## Standard Deviation: 3.466359
##
## Variable: Income
## Mean: 158263.2
## Median: 157866
## Range: 279892
## Standard Deviation: 80575.19
##
## Variable: Physical.Activity.Days.Per.Week
## Mean: 3.489672
## Median: 3
## Range: 7
## Standard Deviation: 2.282687
##
## Variable: Sleep.Hours.Per.Day
## Mean: 7.023508
## Median: 7
## Range: 6
## Standard Deviation: 1.988473

##
## Variable: Sex
##
## Female   Male
##   2652   6111
##
## Variable: Diabetes
##
##    0    1
## 3047 5716
##
## Variable: Family.History
##
##    0    1
## 4443 4320
##
## Variable: Smoking
##
##    0    1
##  904 7859
##
## Variable: Obesity
```

```
##
##    0    1
## 4369 4394
##
## Variable: Alcohol.Consumption
##
##    0    1
## 3522 5241
##
## Variable: Diet
##
##   Average   Healthy Unhealthy
##      2912      2960      2891
##
## Variable: Previous.Heart.Problems
##
##    0    1
## 4418 4345
##
## Variable: Medication.Use
##
##    0    1
## 4396 4367
##
## Variable: Country
##
##      Argentina      Australia         Brazil         Canada          China
##            471            449            462            440            436
##       Colombia         France        Germany          India          Italy
##            429            446            477            412            431
##          Japan    New Zealand        Nigeria   South Africa    South Korea
##            433            435            448            425            409
##          Spain       Thailand United Kingdom  United States        Vietnam
##            430            428            457            420            425
##
## Variable: Continent
##
##        Africa          Asia     Australia        Europe North America
##           873          2543           884          2241           860
## South America
##          1362
##
## Variable: Hemisphere
##
## Northern Hemisphere Southern Hemisphere
##                5660                3103
##
## Variable: Heart.Attack.Risk
```

```
##
##    0    1
## 5624 3139

## Categorical Columns:

## [1] "Patient.ID"     "Sex"            "Blood.Pressure" "Diet"
## [5] "Country"        "Continent"      "Hemisphere"

##
## Numerical Columns:

##  [1] "Age"                              "Cholesterol"
##  [3] "Heart.Rate"                       "Diabetes"
##  [5] "Family.History"                   "Smoking"
##  [7] "Obesity"                          "Alcohol.Consumption"
##  [9] "Exercise.Hours.Per.Week"          "Previous.Heart.Problems"
## [11] "Medication.Use"                   "Stress.Level"
## [13] "Sedentary.Hours.Per.Day"          "Income"
## [15] "BMI"                              "Triglycerides"
## [17] "Physical.Activity.Days.Per.Week" "Sleep.Hours.Per.Day"
## [19] "Heart.Attack.Risk"

##                                                        variable    n      mean
## Age                                                         Age 8763     53.71
## Cholesterol                                         Cholesterol 8763    259.88
## Heart.Rate                                           Heart.Rate 8763     75.02
## Diabetes                                               Diabetes 8763      0.65
## Family.History                                   Family.History 8763      0.49
## Smoking                                                 Smoking 8763      0.90
## Obesity                                                 Obesity 8763      0.50
## Alcohol.Consumption                         Alcohol.Consumption 8763      0.60
## Exercise.Hours.Per.Week                 Exercise.Hours.Per.Week 8763     10.01
## Previous.Heart.Problems                 Previous.Heart.Problems 8763      0.50
## Medication.Use                                   Medication.Use 8763      0.50
## Stress.Level                                       Stress.Level 8763      5.47
## Sedentary.Hours.Per.Day                 Sedentary.Hours.Per.Day 8763      5.99
## Income                                                   Income 8763 158263.18
## BMI                                                         BMI 8763     28.89
## Triglycerides                                     Triglycerides 8763    417.68
## Physical.Activity.Days.Per.Week Physical.Activity.Days.Per.Week 8763      3.49
## Sleep.Hours.Per.Day                         Sleep.Hours.Per.Day 8763      7.02
## Heart.Attack.Risk                             Heart.Attack.Risk 8763      0.36
##                                    median     range
## Age                                 54.00     72.00
## Cholesterol                        259.00    280.00
## Heart.Rate                          75.00     70.00
```

```
## Diabetes                              1.00      1.00
## Family.History                        0.00      1.00
## Smoking                               1.00      1.00
## Obesity                               1.00      1.00
## Alcohol.Consumption                   1.00      1.00
## Exercise.Hours.Per.Week              10.07     20.00
## Previous.Heart.Problems               0.00      1.00
## Medication.Use                        0.00      1.00
## Stress.Level                          5.00      9.00
## Sedentary.Hours.Per.Day               5.93     12.00
## Income                           157866.00 279892.00
## BMI                                  28.77     21.99
## Triglycerides                       417.00    770.00
## Physical.Activity.Days.Per.Week       3.00      7.00
## Sleep.Hours.Per.Day                   7.00      6.00
## Heart.Attack.Risk                     0.00      1.00
```

21 8. Planned statistical methods

As the project progresses, I plan to use chi-square tests to assess associations between categorical factors (e.g., smoking, diabetes) and heart attack risk, and t-tests/ANOVA to compare continuous measures (e.g., cholesterol, BMI) across groups. To build predictive insight, I will apply logistic regression and may explore machine learning models such as decision trees or random forests. These methods will help identify key risk factors and evaluate their predictive power.

22) JOINT PROJECTS - References

Project 1 - Our World in Data. "Coronvirus Pandemic (COVID-19) dataset." source location - Source: URL: https://docs.owid.io/projects/covid/en/latest/dataset.html Project 2 - mclikmb4, (2021, April 4), Coronavirus-dataset France, Kaggle, https://www.kaggle.com/datasets/mclikmb4/coronavirusdataset-france?select=chiffres-cles.csv

Project 3 - Banerjee, S. (2021). Heart Attack Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

23) PROJECTS - Appendix

23.1 Project 1

23.2 Project 2

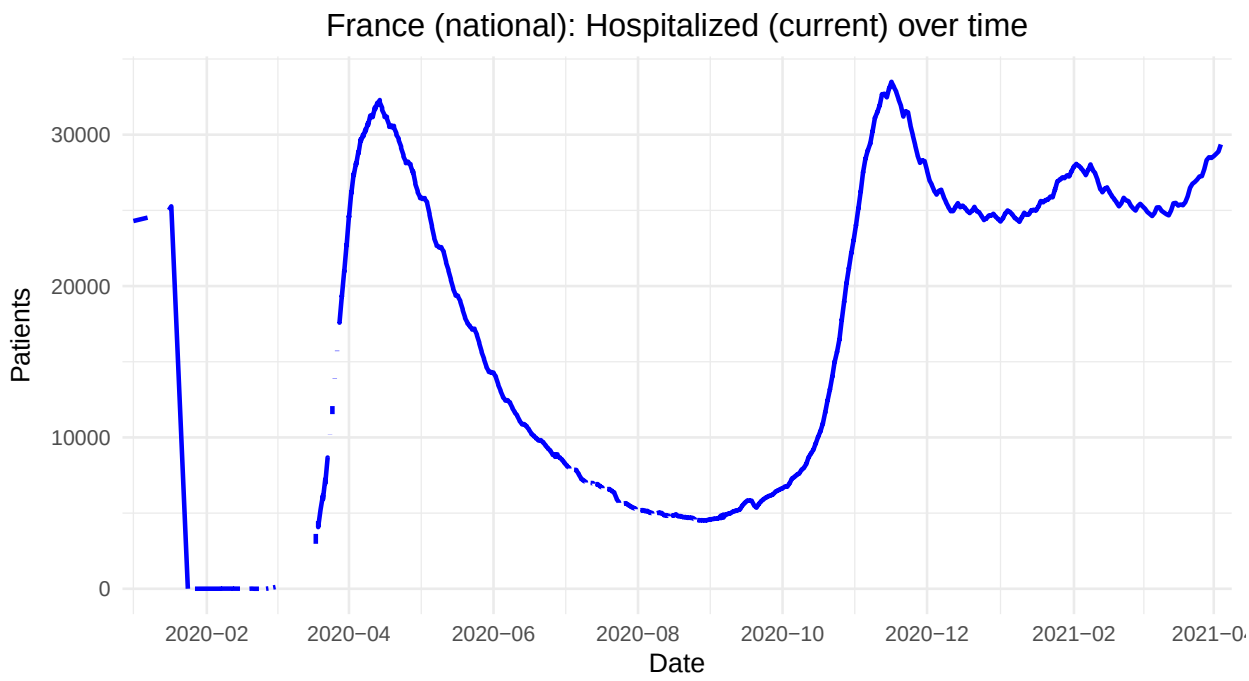


Figure 4: France (national): Hospitalized (current) over time.

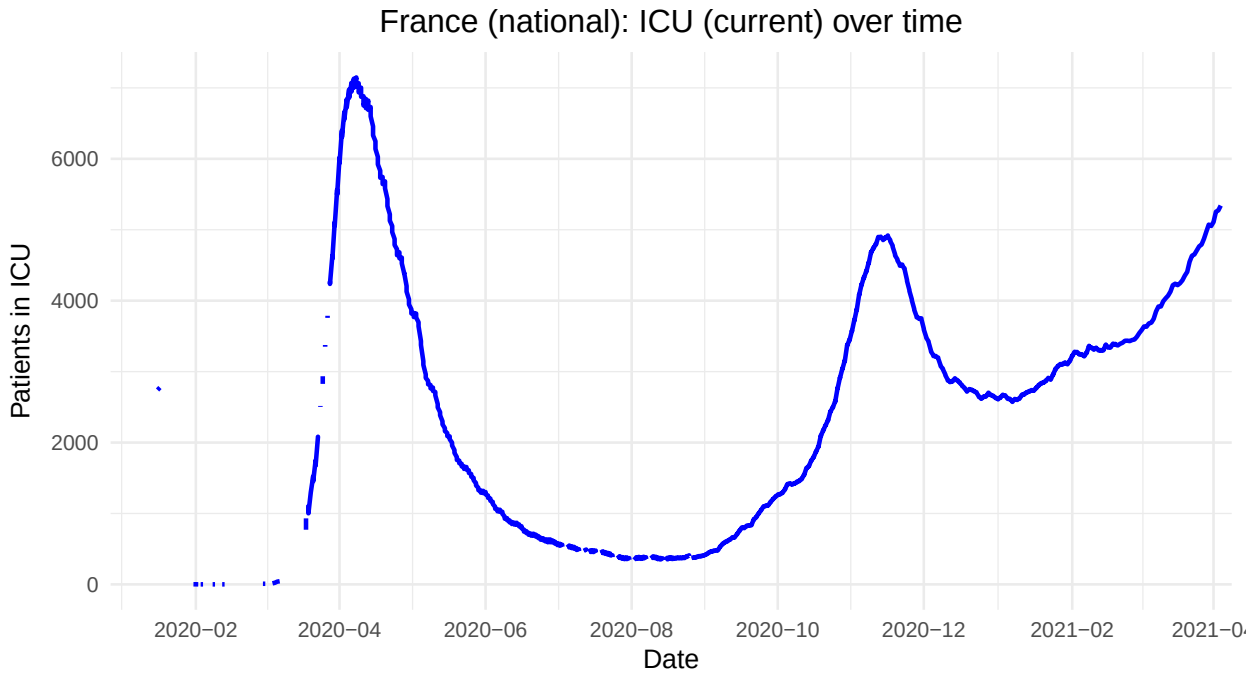


Figure 5: France (national): ICU (current) over time.

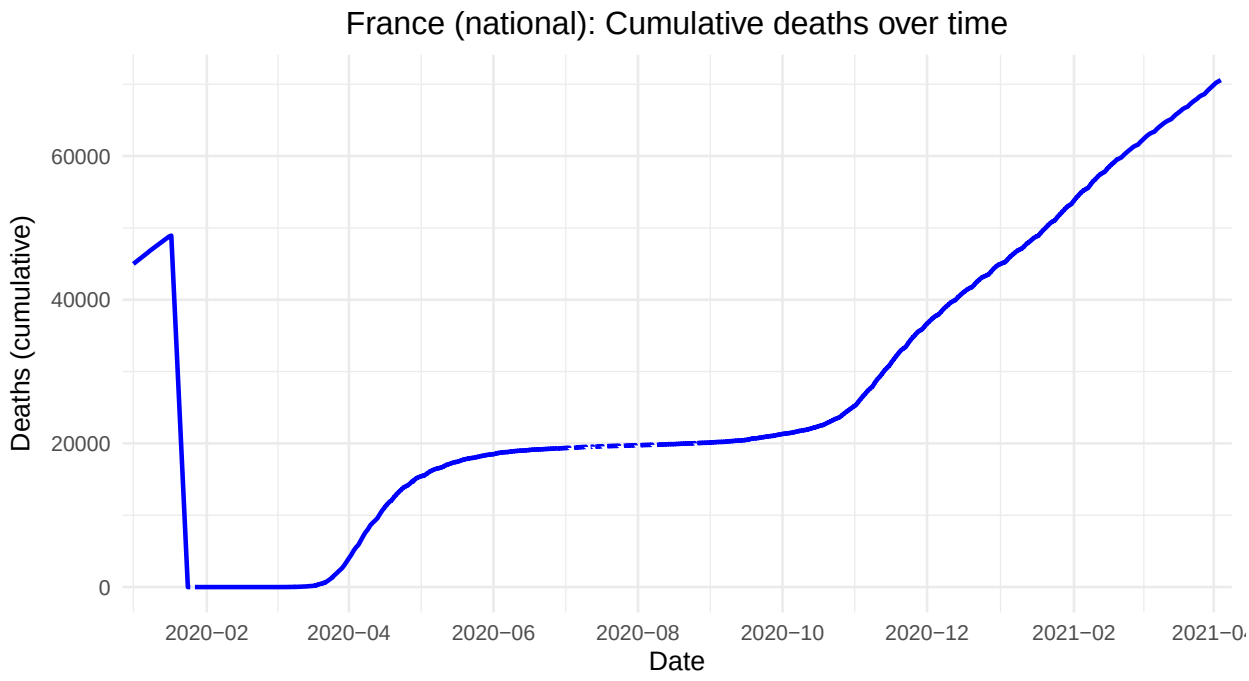


Figure 6: France (national): Cumulative deaths over time.