B518 | Week 4 | Group Project | Submission 1

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

2025-10-01

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

```
## [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"
```

2 Introduction

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.

3 Dataset justification

4 Selection Criteria:

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.

Relevance: Directly biomedical/public-health, reflecting real-world cases and death metrics during COVID-19.

Size/structure: The file far exceeds the minimum requirements (61 columns and 530k rows) and includes both categorical (e.g. Country) and continuous fields (total_deaths_per_million, new_cases_per_million)

Source Location: https://docs.owid.io/projects/covid/en/latest/dataset.html **Raw data Location:** https://catalog.ourworldindata.org/garden/covid/latest/compact/compact.csv

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

Analytical potential: Feasible for tables, histograms, boxplots, time trends. Using the fields with the suffix "per_million" allows better scaling for cross country comparisons and summaries.

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.

5 Variables and structure

This analysis focuses on a few key variables from the dataset. The primary categorical variable is 'country', which we have filtered to four specific nations. The main continuous variable is 'new_cases_per_million', which allows for a fair comparison of infection rates by account for population differences. Finally, the 'date' variable was used to filter the data to the 2021 calendar year.

A list of all the fields: - "country" - "date" - "total_cases", "total_cases_per_million" - "new_cases", "new_cases_smoothed", "new_cases_per_million", "new_cases_smoothed_per_million" - "total_deaths", "new_deaths", "new_deaths_smoothed", "total_deaths_per_million"

6 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

7 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.

8 Descriptive statistics (figures in Appendix)

Our descriptive analysis for 2021 reveals starkly different pandemic experiences among the four selected countries. The most significant finding is the extreme contrast between China, which reported virtually no community spread, and the western nations all expereicned substantial waves of the Covid virus.

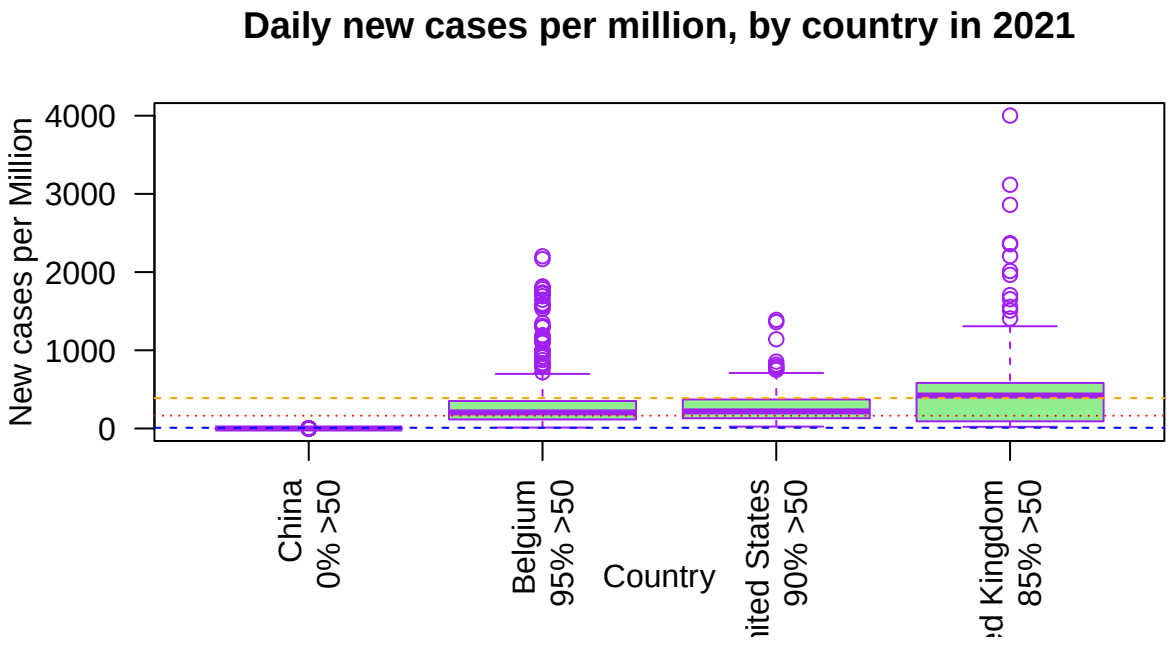
Answering our first research question, the two way frequency shows that China had reported zero days exceeding a 50 new cases per million threshold. In contrast this threshold was crossed 94.% of days in Belgium, 89.9% in the USA and 84.9% in the UK.

For our second question, the summary statistics table identified the UK as having the highesy typical daily caseload, with a median of 421.6 new cases per million, nearly double that of the USA at 218.4. The overall distrubution o cases is heavily right skewed, a pattern confirmed visually by the histogram and the numerous high end outliers visibl in the boxplot.

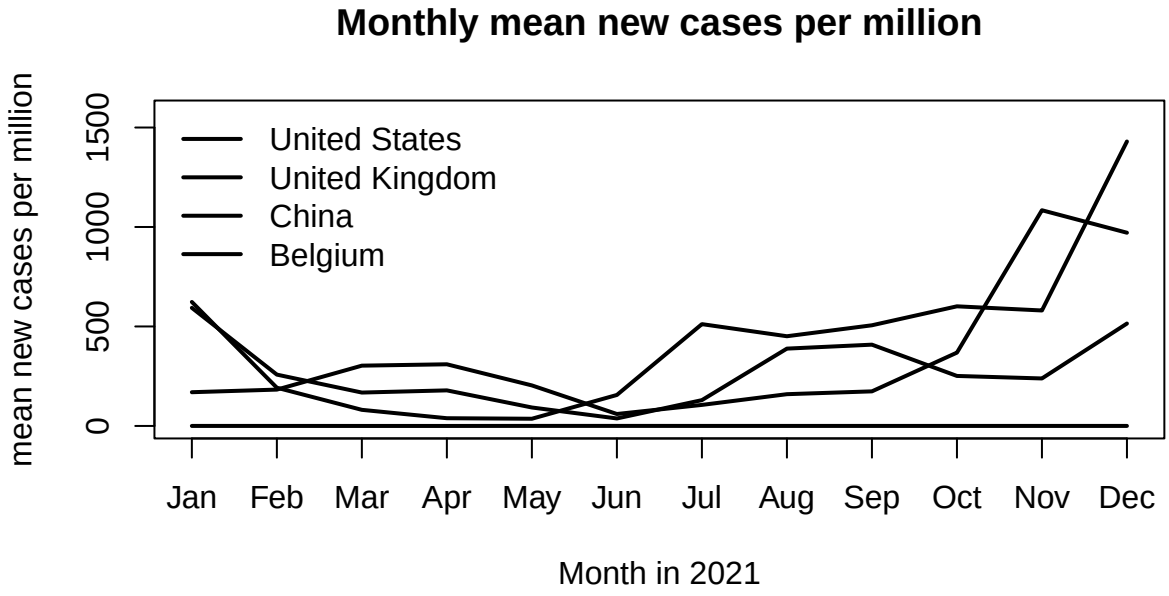
Lastly, for our third research question. the time trend plot issustrates how these case rates evolved monthly. China's rate remained flat, the US, UK and Belgium all expereicned a summer drop followed by a big surge in Q4 of 2021.

8.1 Boxplot (numeric by category)

Boxplot (Mortality rate by category)



8.2 Time trend (average by year)



9 Planned statisical methods

To formally test for differences in the median daily new cases between the four countries, we plan to use a non-parametric test such as the Kruskal-Wallis test, given the skewed nature of the data. Further analysis could involve using correlation to explore the relationship between vaccination rates and new cases over time for each country.

10 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

11 Appendix - Project One

11.1 One-Way frequency table (categorical)

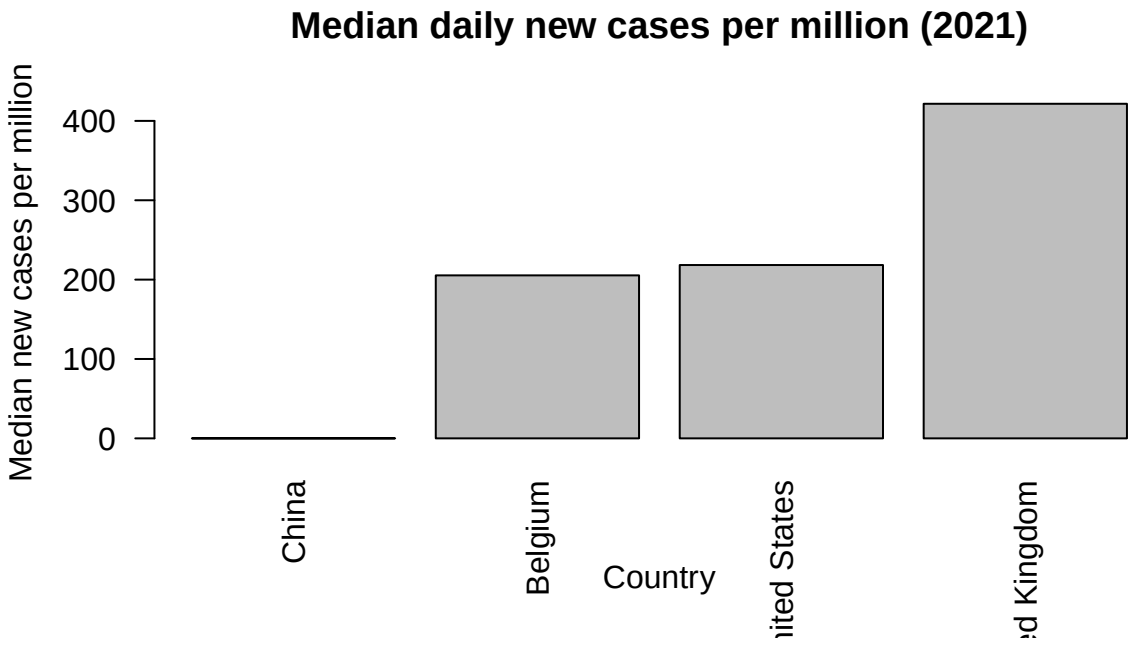
Counts and proportions for a categorical variable

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

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

11.2 Bar Chart of Disease.Category (counts)

Bar chart / Bar plot of disease category by count



11.3 Two way table (category by category)

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

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

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

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

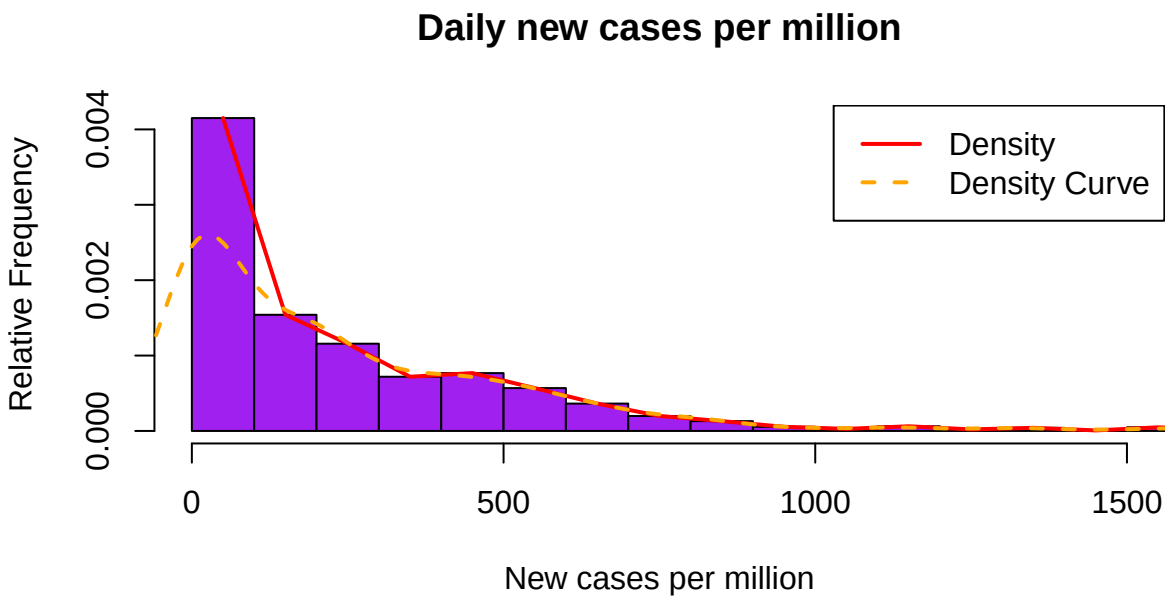
```
##          China        Belgium  United States United Kingdom
##          0.000          0.945          0.899          0.849
```

11.4 Center & Spread (overall, selected countries, 2021)

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

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

11.5 Histogram (shape of the distrubution)



12 Project Idea Two - Covid 19 Hospitalizations in France

13 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

14 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).

15 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.

16 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
tested	0	NaN	NA	NA	Inf	-Inf

17 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?

18 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.

19 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.

20 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).

21 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.

22 Appendix - Project Two

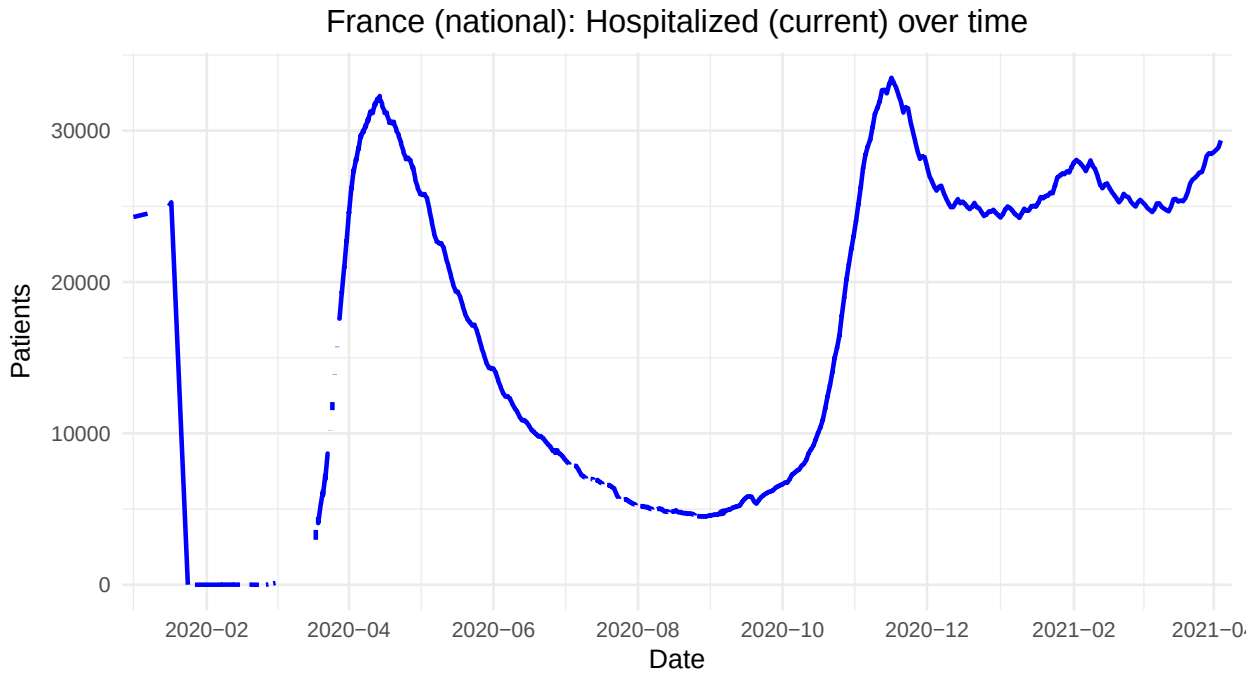


Figure 1: France (national): Hospitalized (current) over time.

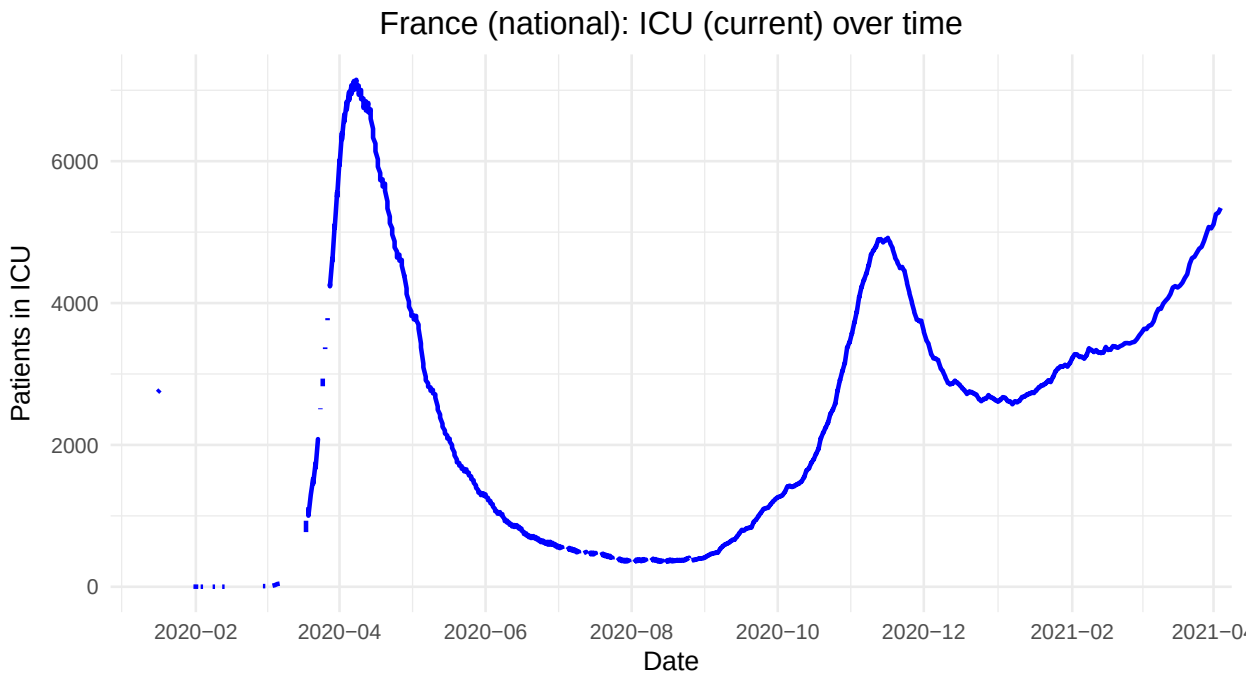


Figure 2: France (national): ICU (current) over time.

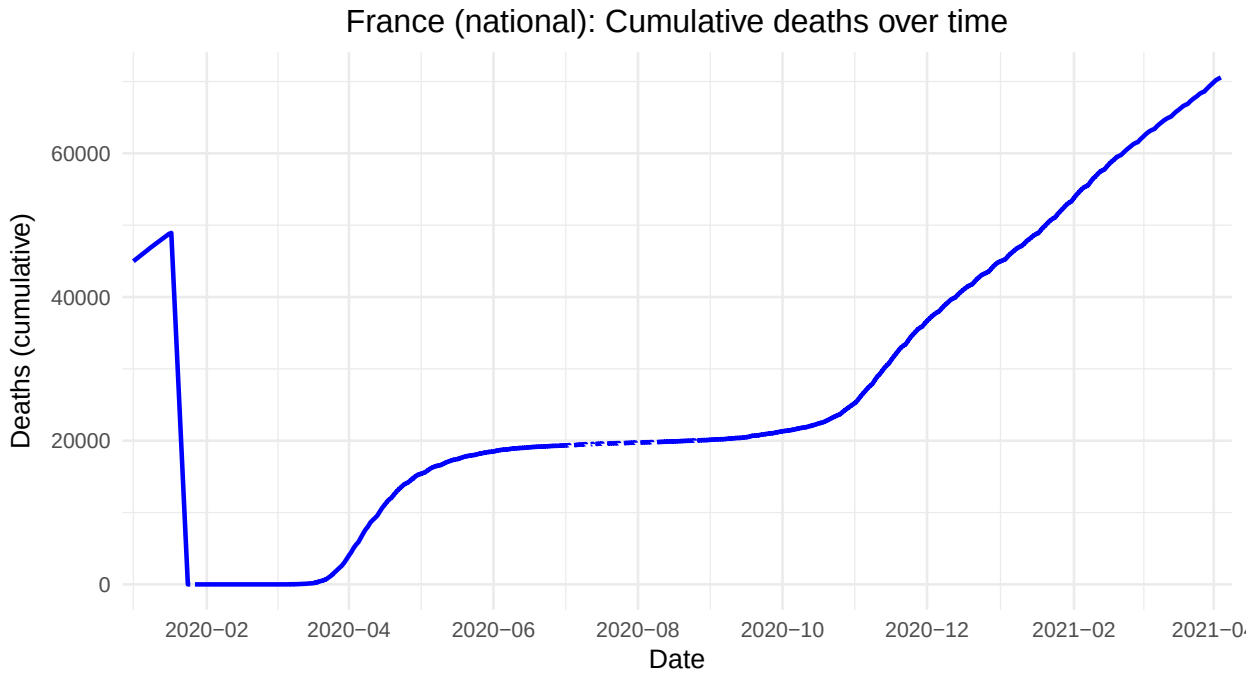


Figure 3: France (national): Cumulative deaths over time.

23 Project Idea Three - Heart attack

24 Link to the dataset

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

25 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.

26 Dataset justification

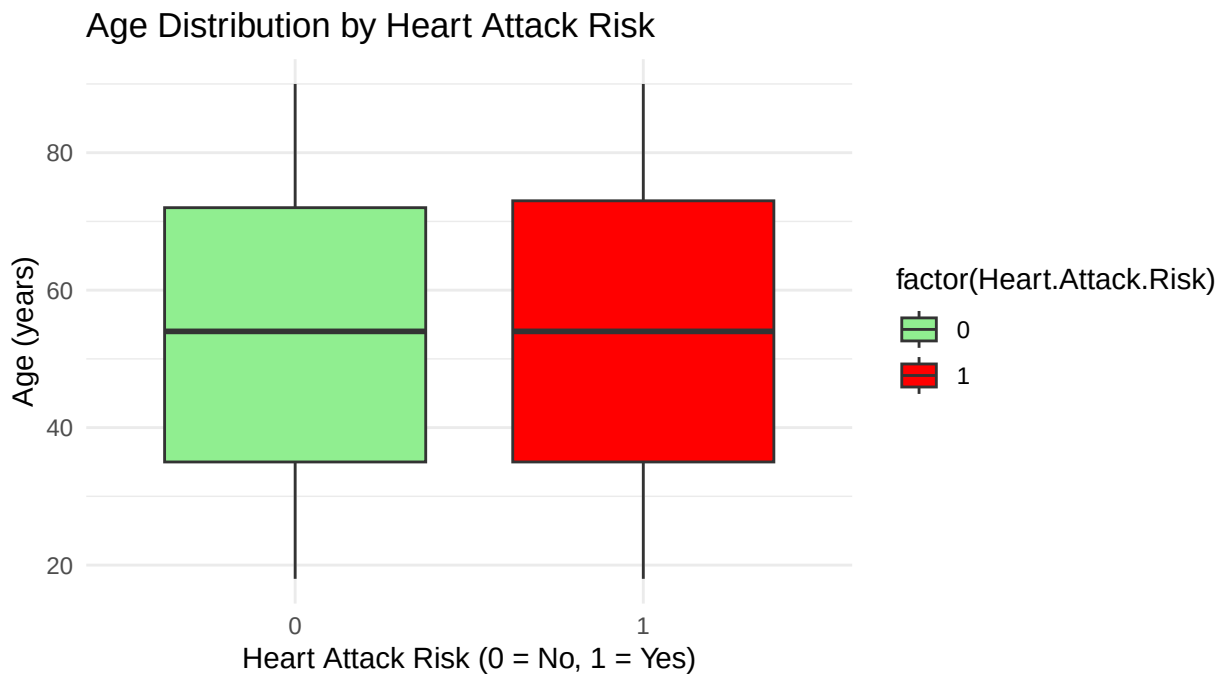
I chose the Heart Attack Prediction Dataset because it directly addresses a critical biomedical challenge cardiovascular disease which remains one of the leading causes of mortality worldwide. The dataset integrates clinical, lifestyle, and demographic variables, making it highly relevant for exploring 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.

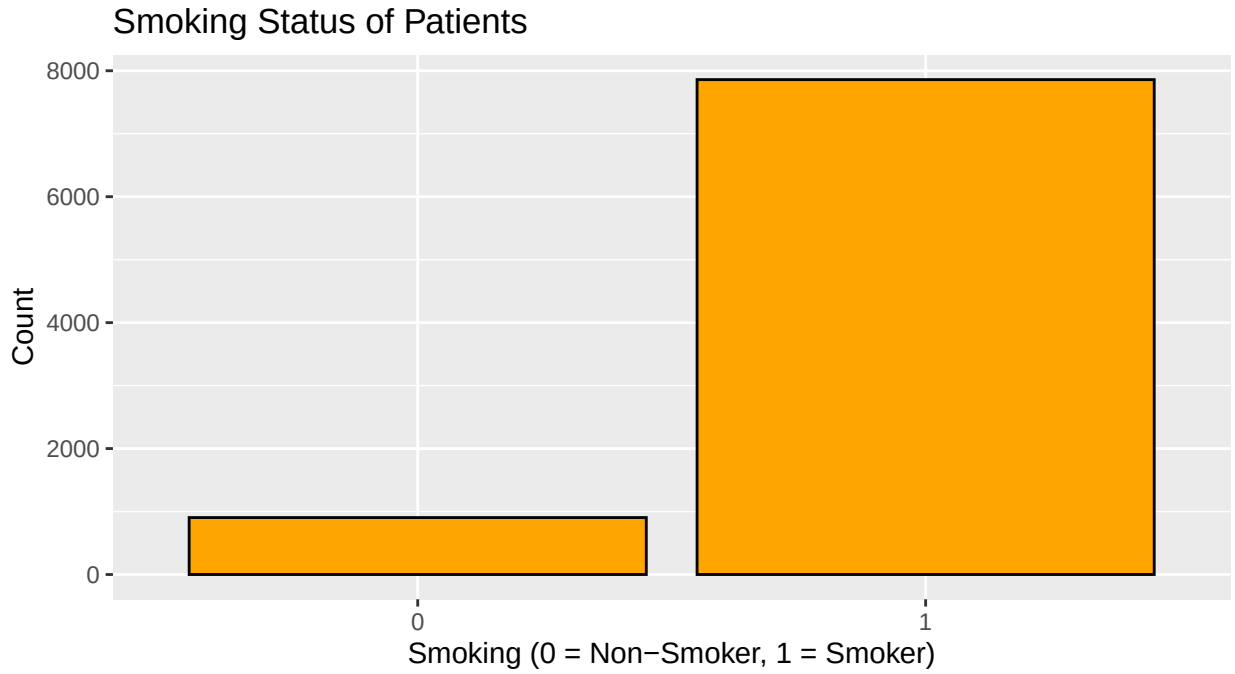
27 Variables description

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
```





28 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?

29 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 over-representation 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.

30 Descriptive statistics and data visualizations

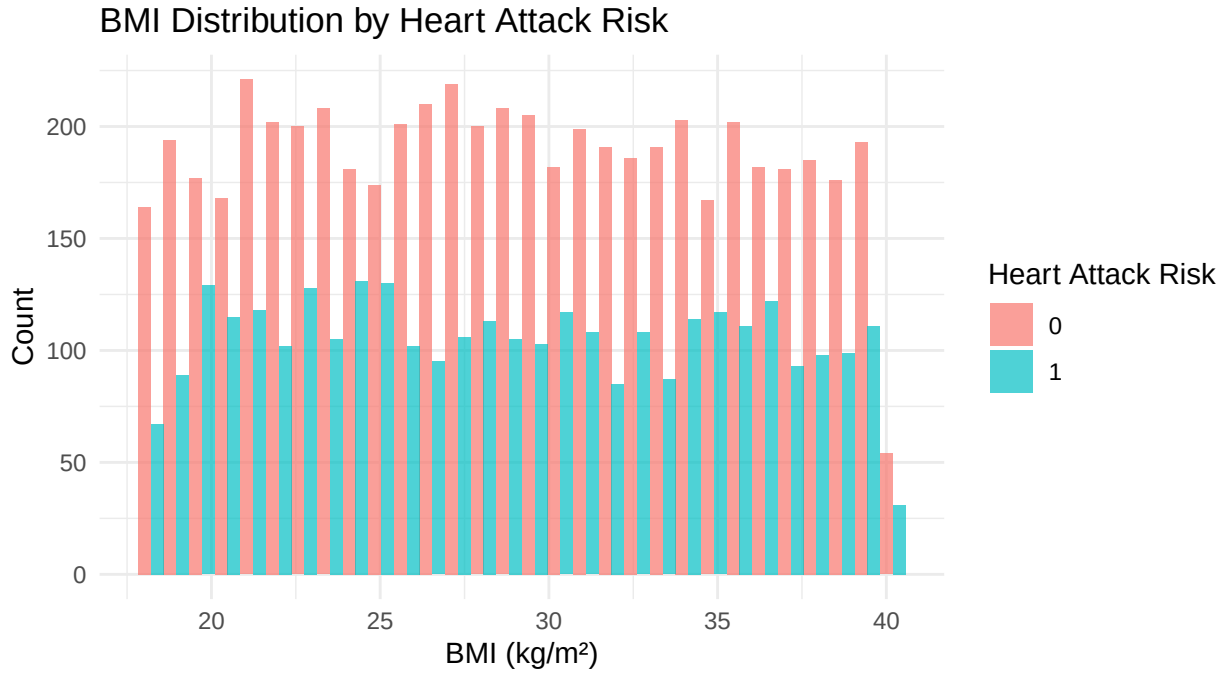
```
##
## Variable: Age
## Mean: 53.70798
## Median: 54
## Range: 72
## Standard Deviation: 21.24951
##
## Variable: Cholesterol
## Mean: 259.8772
## Median: 259
## Range: 280
## Standard Deviation: 80.86328
##
## Variable: Heart.Rate
## Mean: 75.02168
## Median: 75
## Range: 70
## Standard Deviation: 20.55095
##
## Variable: BMI
## Mean: 28.89145
## Median: 28.769
```

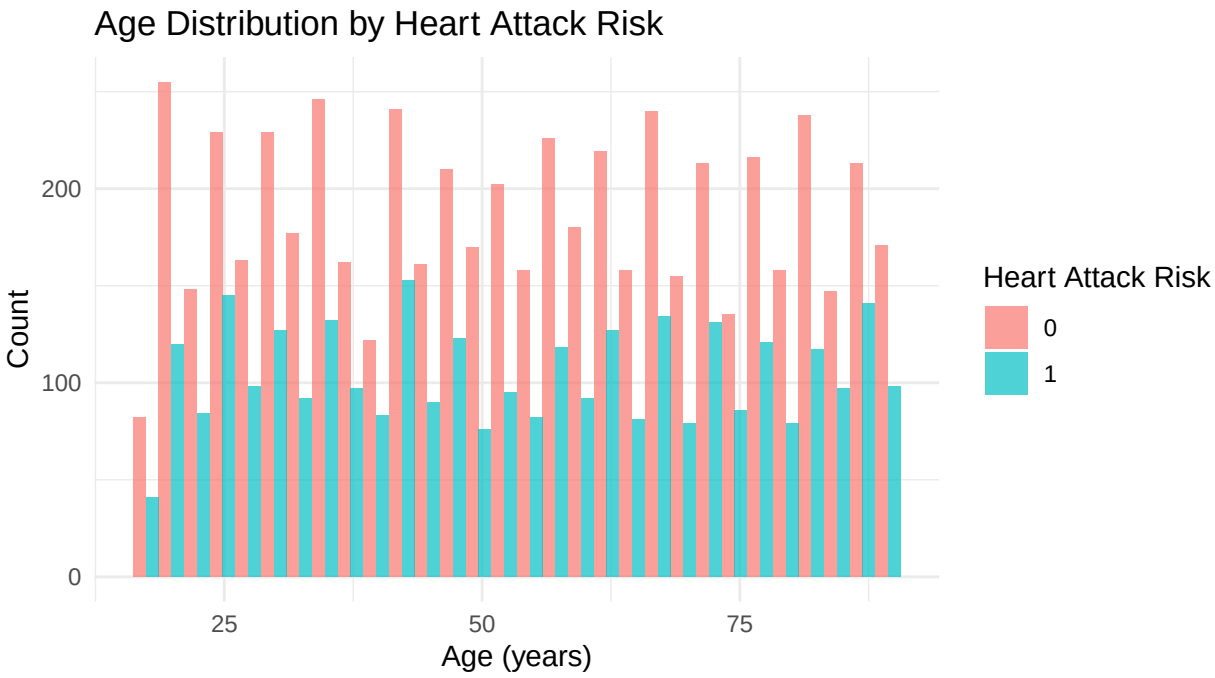
```
## Range: 21.99487
## Standard Deviation: 6.319181
##
## Variable: Triglycerides
## Mean: 417.6771
## Median: 417
## Range: 770
## Standard Deviation: 223.7481
##
## Variable: Exercise.Hours.Per.Week
## Mean: 10.01428
## Median: 10.06956
## Range: 19.99627
## Standard Deviation: 5.783745
##
## Variable: 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
##
```

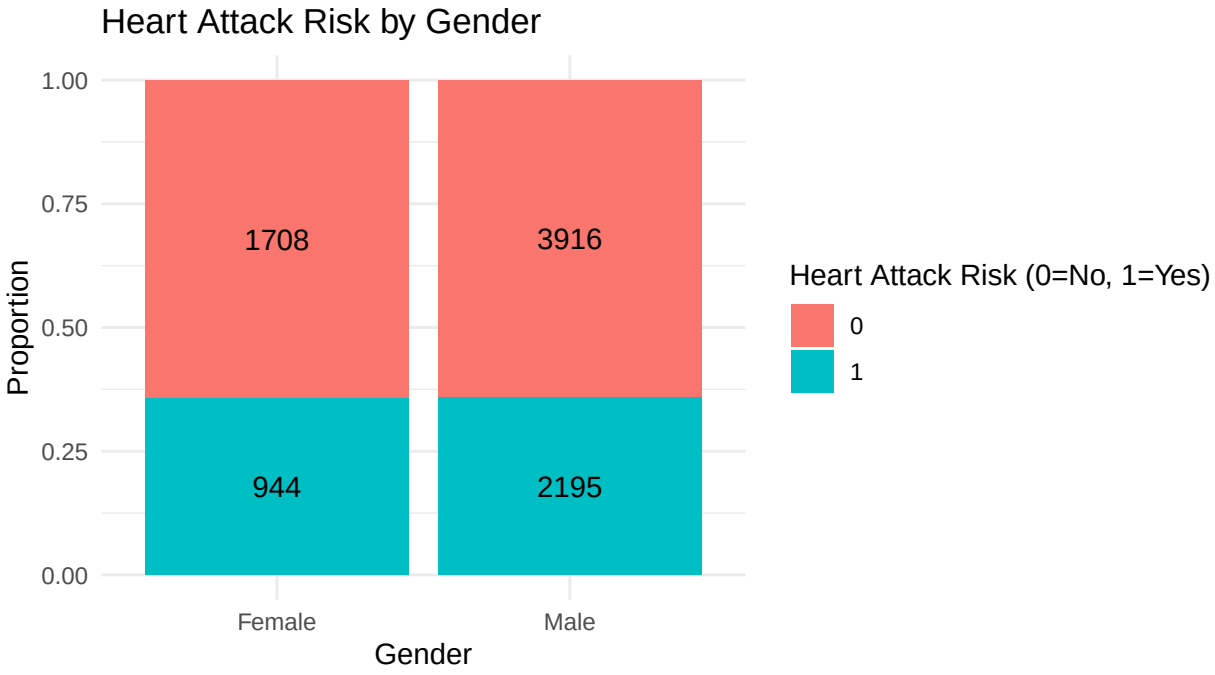
```
## Female   Male
##   2652   6111
##
## Variable: Diabetes
##
##    0    1
## 3047 5716
##
## Variable: Family.History
##
##    0    1
## 4443 4320
##
## Variable: Smoking
##
##    0    1
##  904 7859
##
## Variable: Obesity
##
##    0    1
## 4369 4394
##
## Variable: Alcohol.Consumption
##
##    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
```

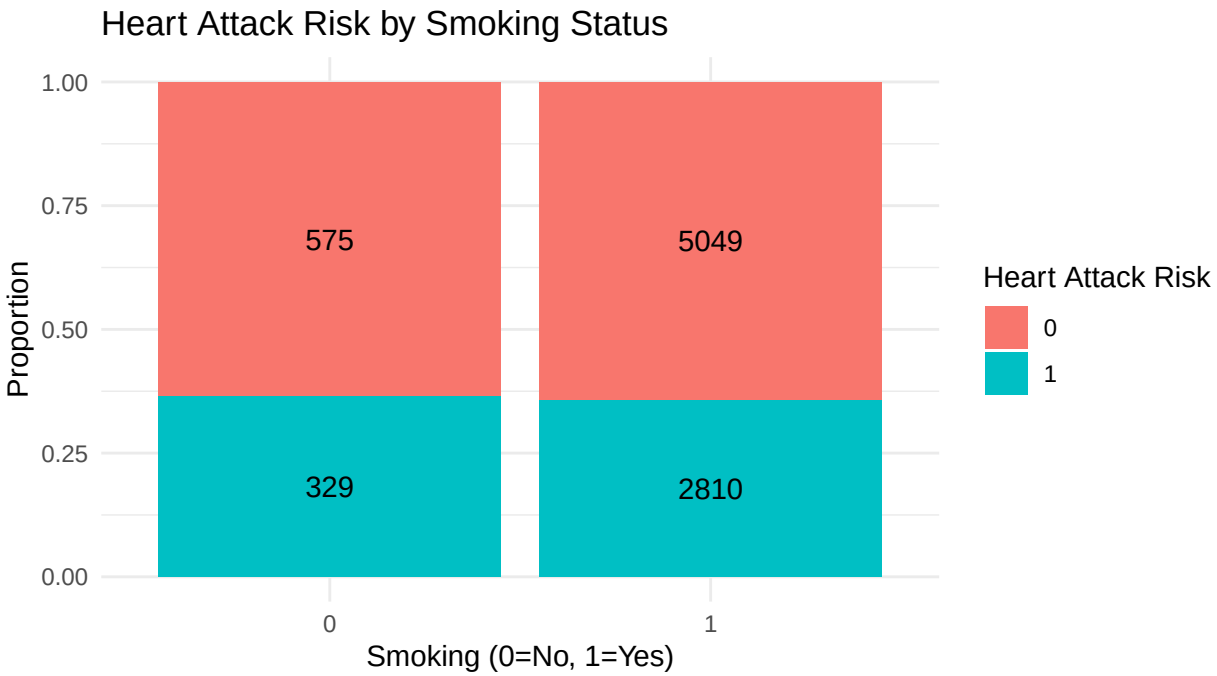
```
##             429             446             477             412             431
##           Japan     New Zealand         Nigeria    South Africa     South Korea
##             433             435             448             425             409
##           Spain        Thailand  United Kingdom   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
```

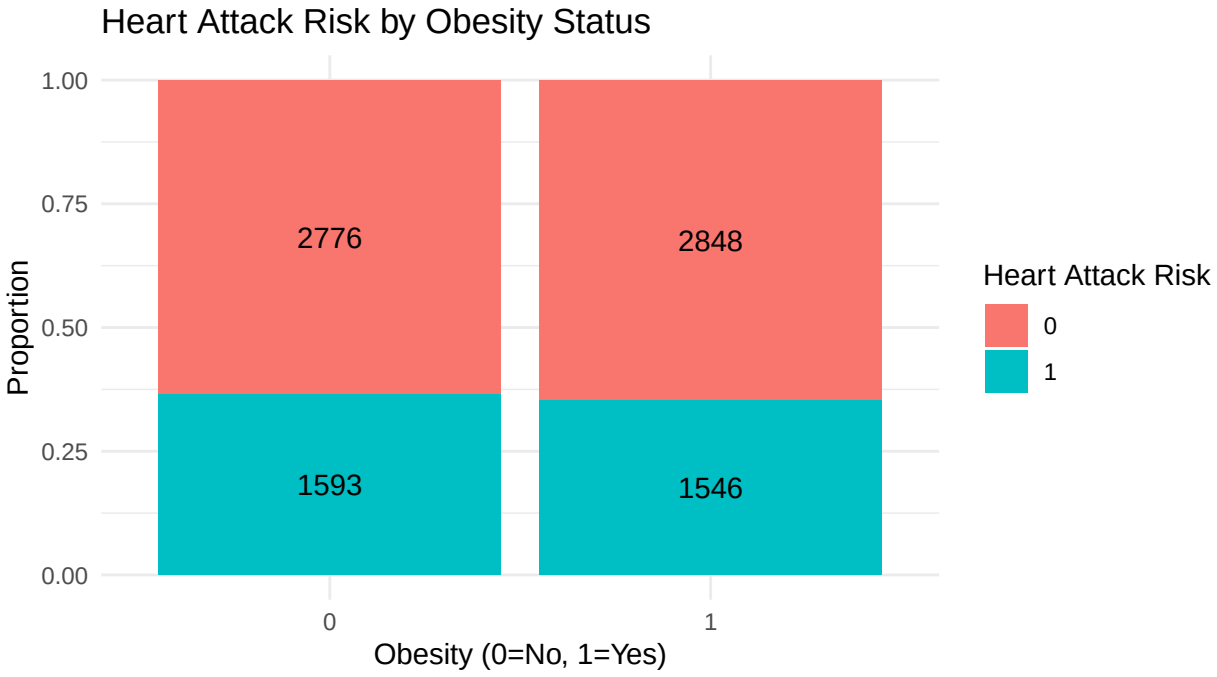
31 Distribution of Heart Attack Risk Across Demographic and Lifestyle Factors

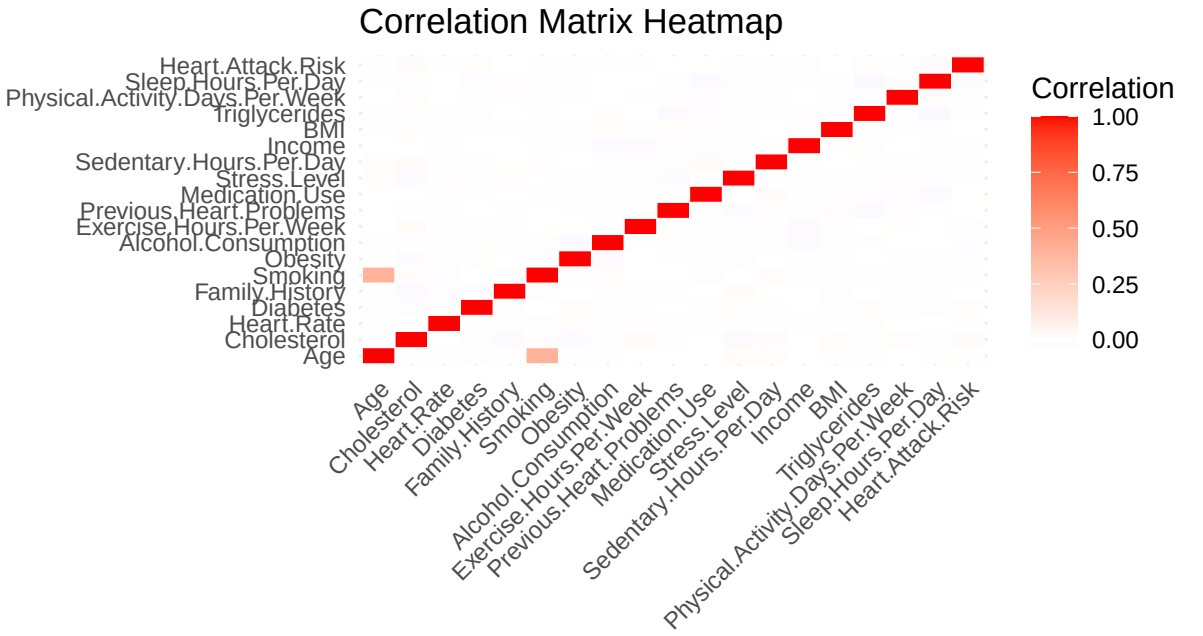



Age Distribution by Heart Attack Risk
Count
200
100
0
25
50
75
Age (years)
Heart Attack Risk
0
1


Heart Attack Risk by Gender
Proportion
1.00
0.75
0.50
0.25
0.00
1708
3916
944
2195
Female
Male
Gender
Heart Attack Risk (0=No, 1=Yes)
0
1


Heart Attack Risk by Smoking Status
1.00
0.75
0.50
0.25
0.00
Proportion
575
5049
329
2810
0
1
Smoking (0=No, 1=Yes)
Heart Attack Risk
0
1


Heart Attack Risk by Obesity Status
1.00
0.75
0.50
0.25
0.00
Proportion
2776
2848
1593
1546
0
1
Obesity (0=No, 1=Yes)
Heart Attack Risk
0
1



32 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.

33 Limitations

While descriptive statistics provide valuable insights into the dataset, they also have some limitations. Measures like mean and standard deviation can be influenced by outliers, which may distort the true central tendency and variability of the data. Categorical variables summarized with frequency counts may oversimplify complex health behaviors, such as smoking or alcohol consumption, without capturing intensity or duration. The dataset itself may contain missing values, inconsistencies, or biases due to self-reported measures (e.g., diet, stress level, or exercise). Additionally, descriptive statistics do not establish causal relationships; they only describe patterns. Therefore, more advanced statistical methods and inferential analyses are needed to draw meaningful conclusions about risk factors for heart attacks.

34 Appendix - Project Three

35) JOINT PROJECTS - References

- Project 1 - Our World in Data. (2024). Coronvirus Pandemic (COVID-19) dataset. 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