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)

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

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.

4 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"

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

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.

7 Descriptive statistics (figures in Appendix)

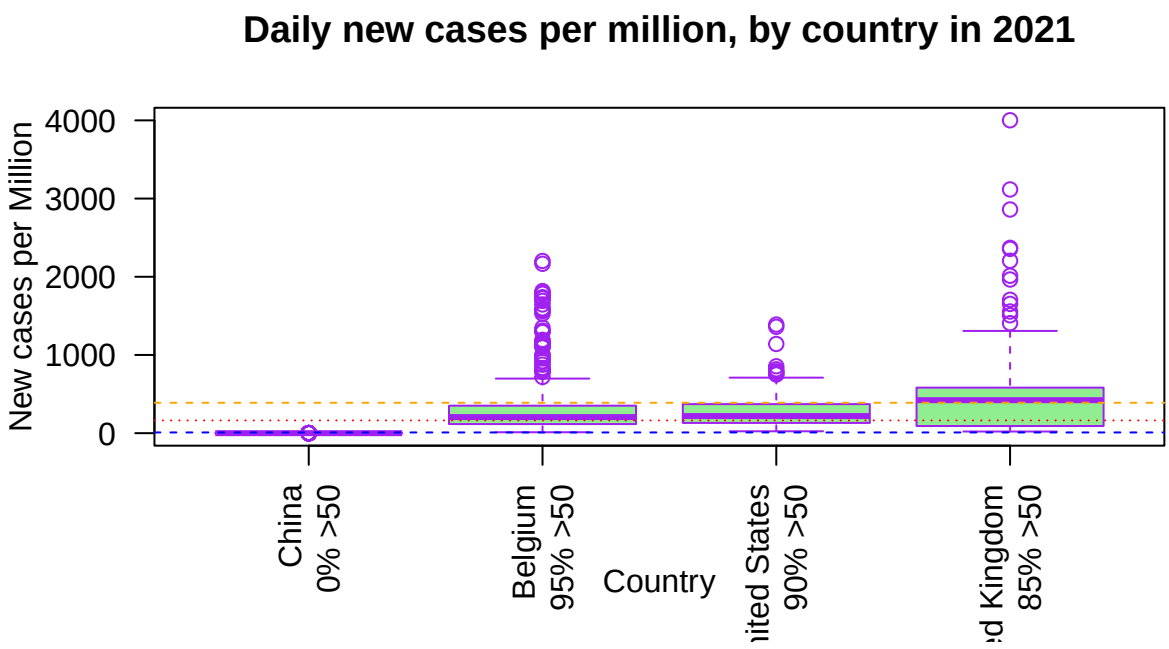
Our descriptive analysis for 2021 reveals starkly different pandemic experiences among the four selected countries. The most significant finding is the extreme contrast between China, which reported virtually no community spread, and the western nations all expereicned substantial waves of the Covid virus. Answering our first research question, the two way frequency (see appendix) 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 (see appendix) 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 (see appendix) and the numerous high end outliers visible in the boxplot.

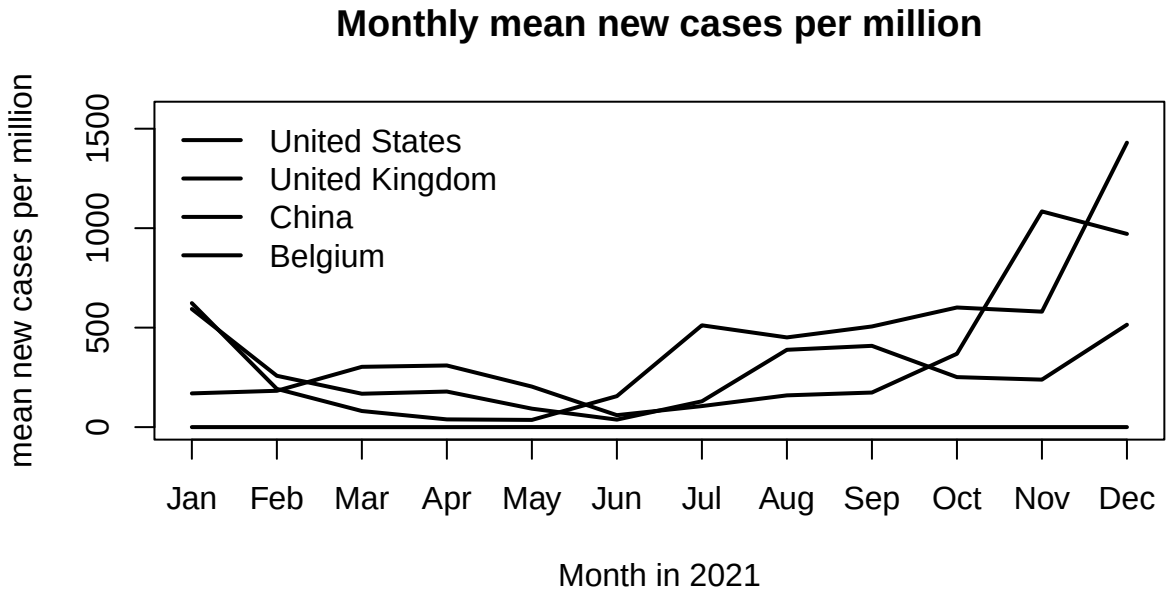
Lastly, for our third research question. the time trend plot issustrates how these case rates evolved monthly. China's rate remained flat, the US, UK and Belgium all expereicned a summer drop followed by a big surge in Q4 of 2021.

7.1 Boxplot (numeric by category)

Boxplot (Mortality rate by category)



7.2 Time trend (average by year)



8 Planned statisical methods

Given the skewed distributions of daily new cases, we will compare medians and IQR's across the four countries and visuaise differences with boxplots and monthly time-trend plots. We will defer formal hypothesis testing to later. However, with that said we may consider a Krushkal-Wallis test to compare medians across countries and examine the relationship between baccination measures and new cases after and if we felt the need to expand the dataset in order to include vaccination variables.

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

10 Appendix - Project One

10.1 One-Way frequency table (categorical)

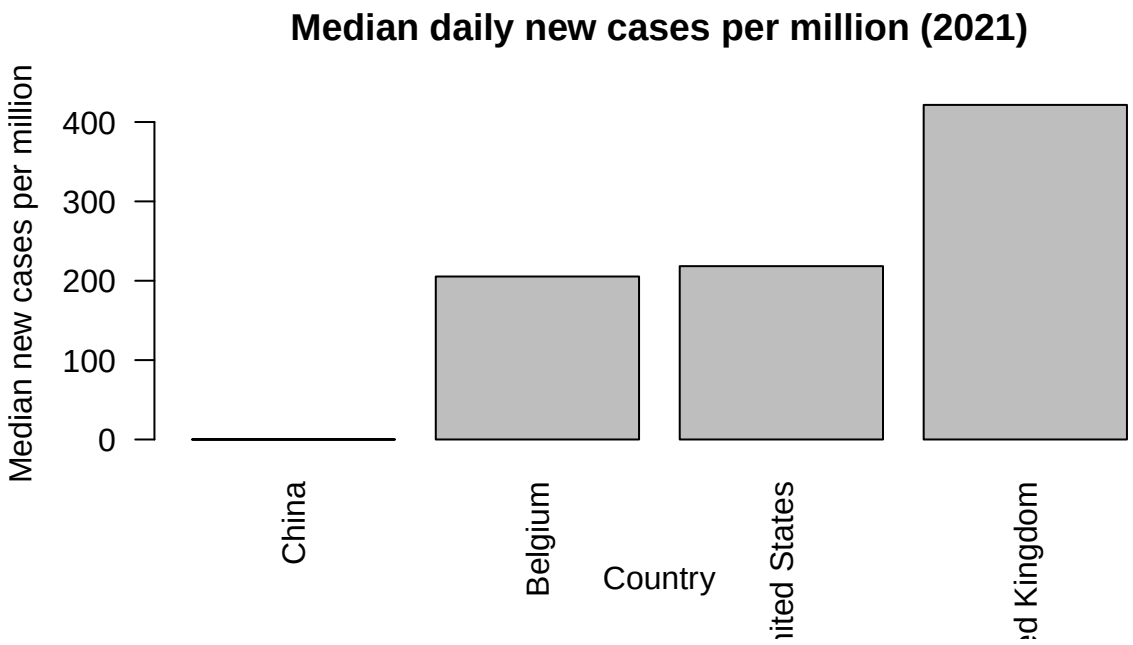
Counts and proportions for a categorical variable

Table 1: One-way table of Country: counts and proportions

country	count	proportion
China	365	0.25
Belgium	365	0.25
United States	365	0.25
United Kingdom	365	0.25

10.2 Bar Chart of Disease.Category (counts)

Bar chart / Bar plot of disease category by count



10.3 Two way table (category by category)

Table 2: Two-way table: Country × High-day indicator (counts)

	Low/Normal	High
China	365	0
Belgium	20	345
United States	37	328
United Kingdom	55	310

Table 3: Row proportions: P(High/Low | Country)

	Low/Normal	High
China	1.000	0.000
Belgium	0.055	0.945
United States	0.101	0.899
United Kingdom	0.151	0.849

Table 4: Column proportions: P(Country | High/Low)

	Low/Normal	High
China	0.765	0.000
Belgium	0.042	0.351
United States	0.078	0.334
United Kingdom	0.115	0.315

Table 5: Two-way table with margins (counts)

	Low/Normal	High	Sum
China	365	0	365
Belgium	20	345	365
United States	37	328	365
United Kingdom	55	310	365
Sum	477	983	1460

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

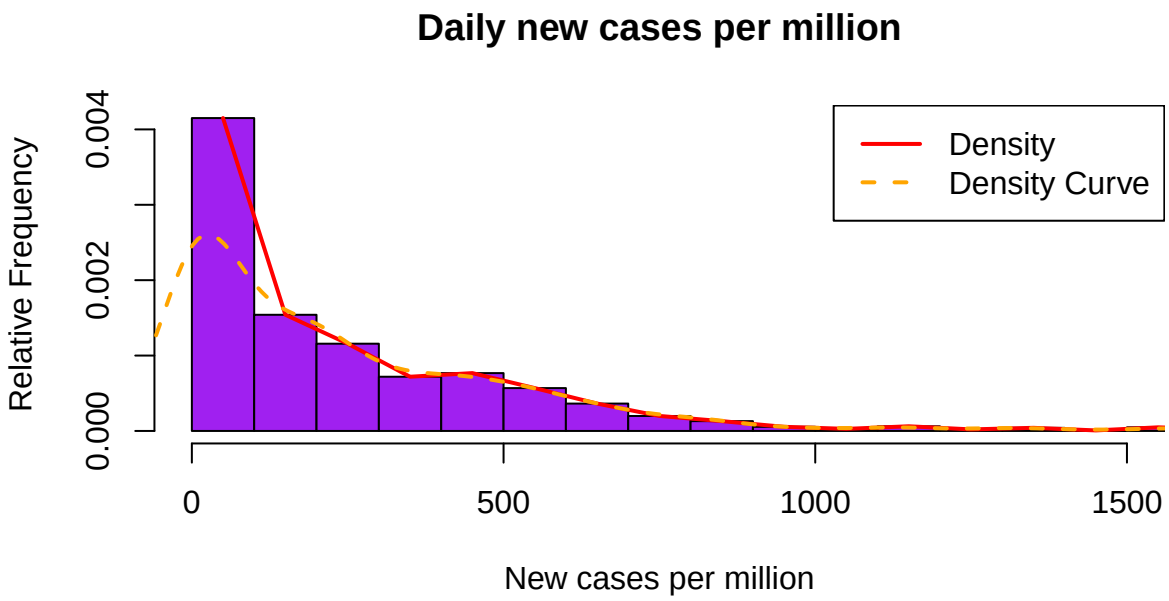
Table 6: Overall center & spread (new_cases_per_million)

median	IQR	sd
162.9	379.7	357.4

Table 7: By-country center & spread (new_cases_per_million)

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

10.5 Histogram (shape of the distrubution)



11 Project Idea Two - Covid 19 Hospitalizations in France

12 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

13 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 the uneven distribution. These indicators and their definitions are documented on the Kaggle dataset page (mclikmb4, 2020-2021).

14 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.

15 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 8: Row counts by geographic granularity

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

Table 9: 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

16 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?

17 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.

18 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.

19 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).

20 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.

21 Appendix - Project Two

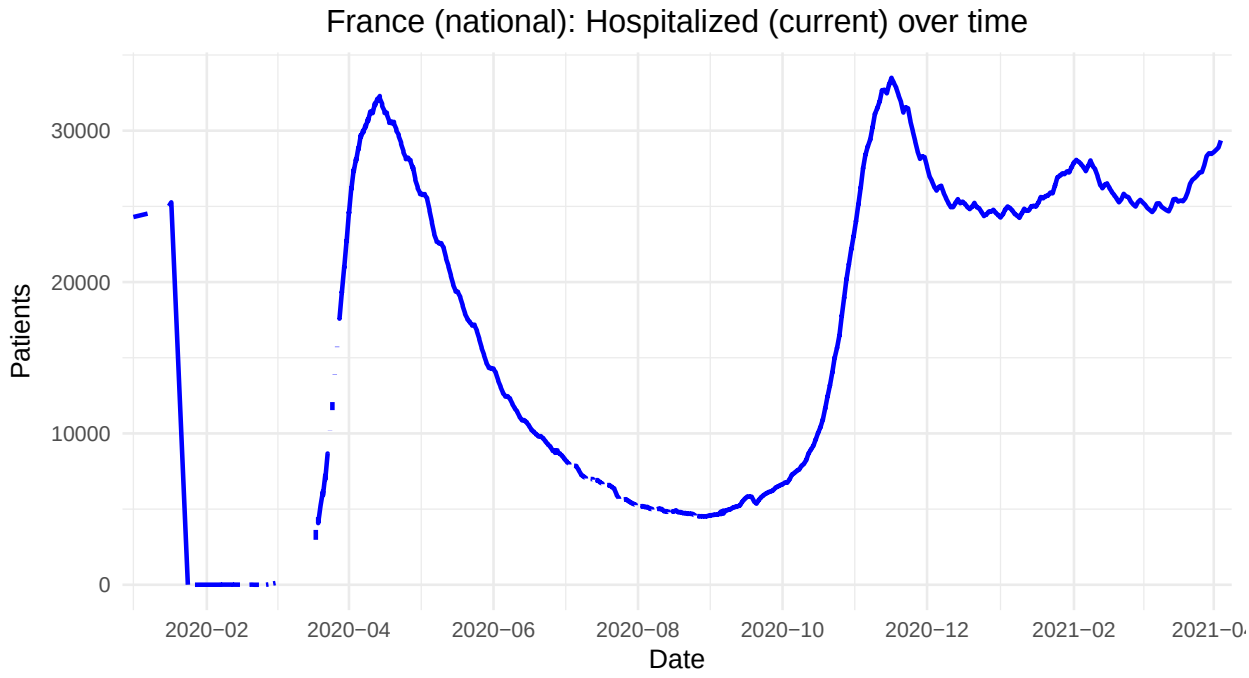


Figure 1: France (national): Hospitalized (current) over time.

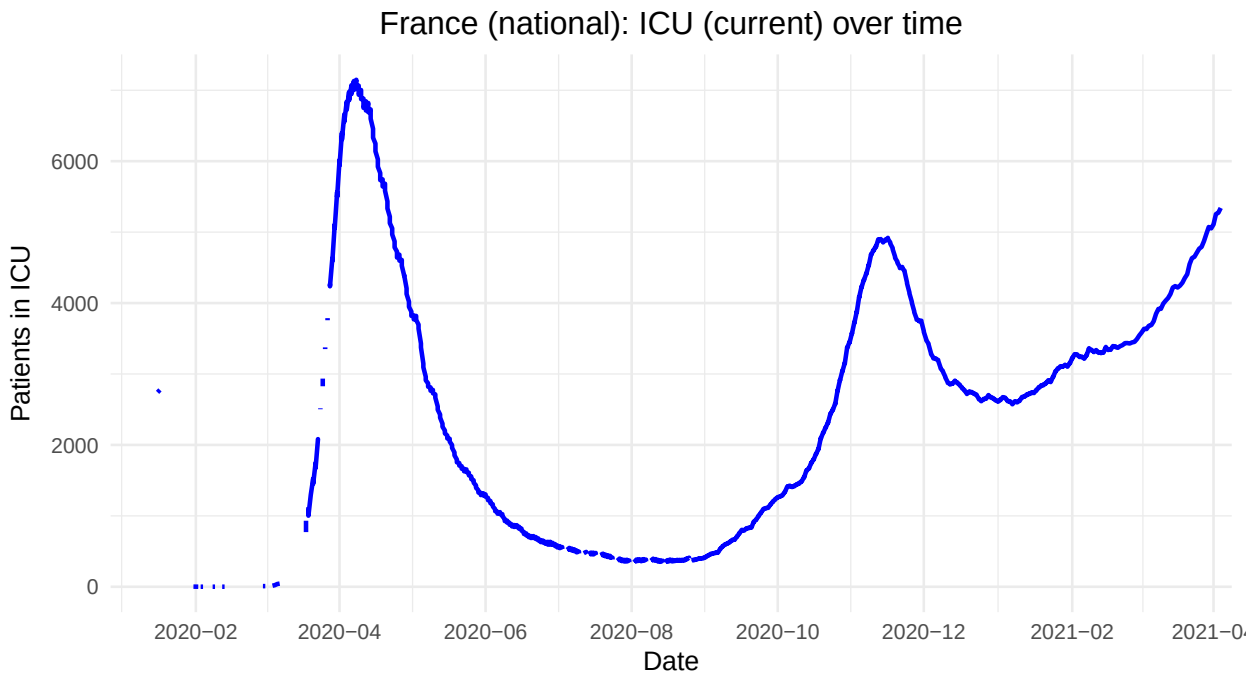


Figure 2: France (national): ICU (current) over time.

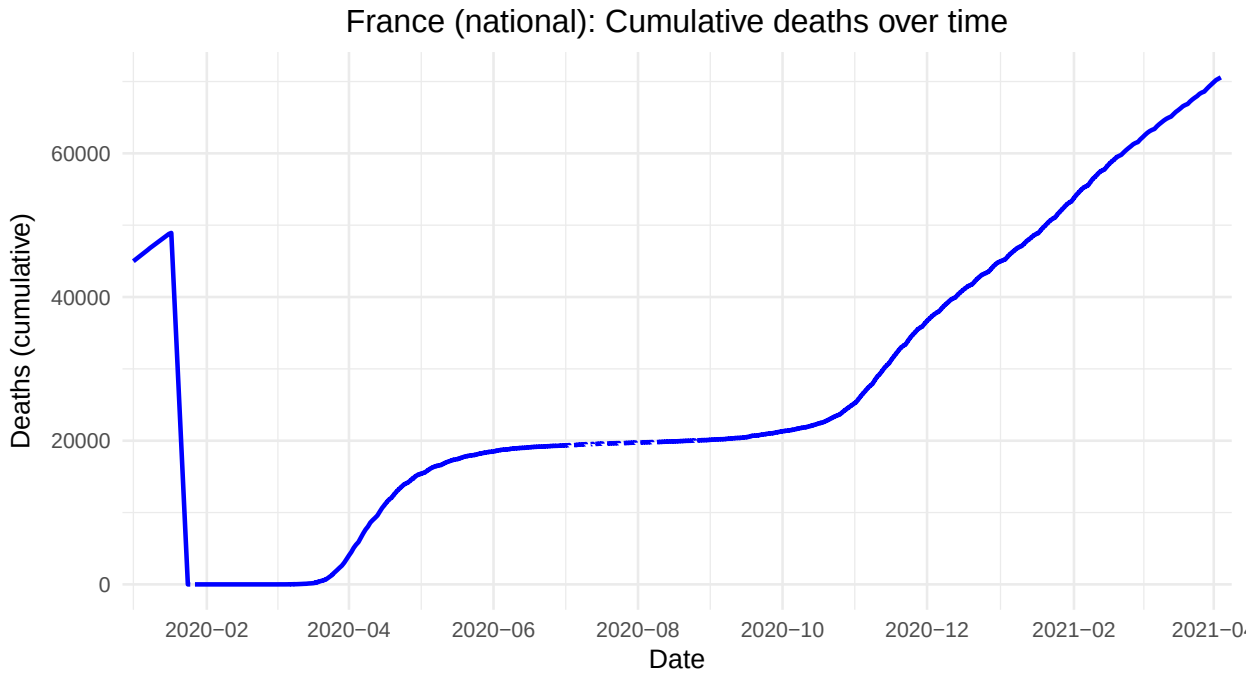


Figure 3: France (national): Cumulative deaths over time.

22 Project Idea Three - Heart attack

23 Link to the dataset

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

24 Introduction to dataset

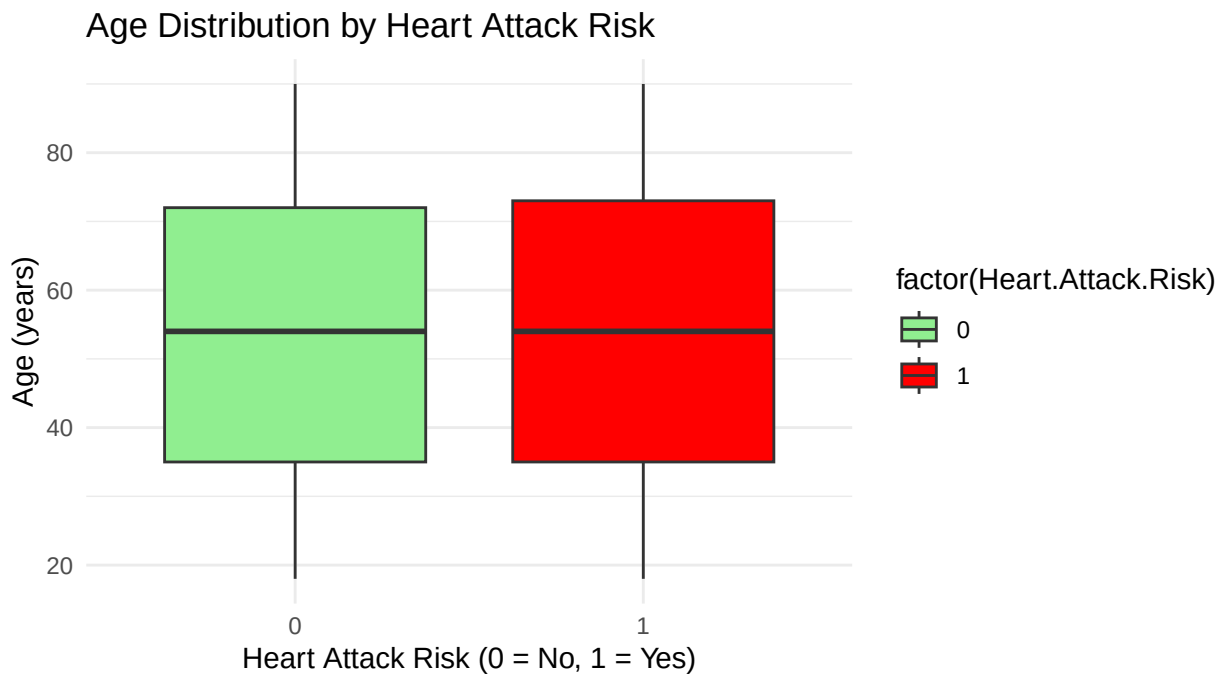
The Heart Attack Prediction Dataset, available on Kaggle, is a comprehensive resource for studying the clinical, lifestyle, and demographic factors associated with cardiovascular risk. It consists of 8,763 de-identified patient records, including continuous variables such as age, cholesterol, blood pressure, 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, help to enhance the dataset by adding more 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 this project. It has a large mix of variables that support exploration of correlations, risk factors, and group comparisons, while also providing an ethical and accessible foundation for predictive modeling in cardiovascular health research.

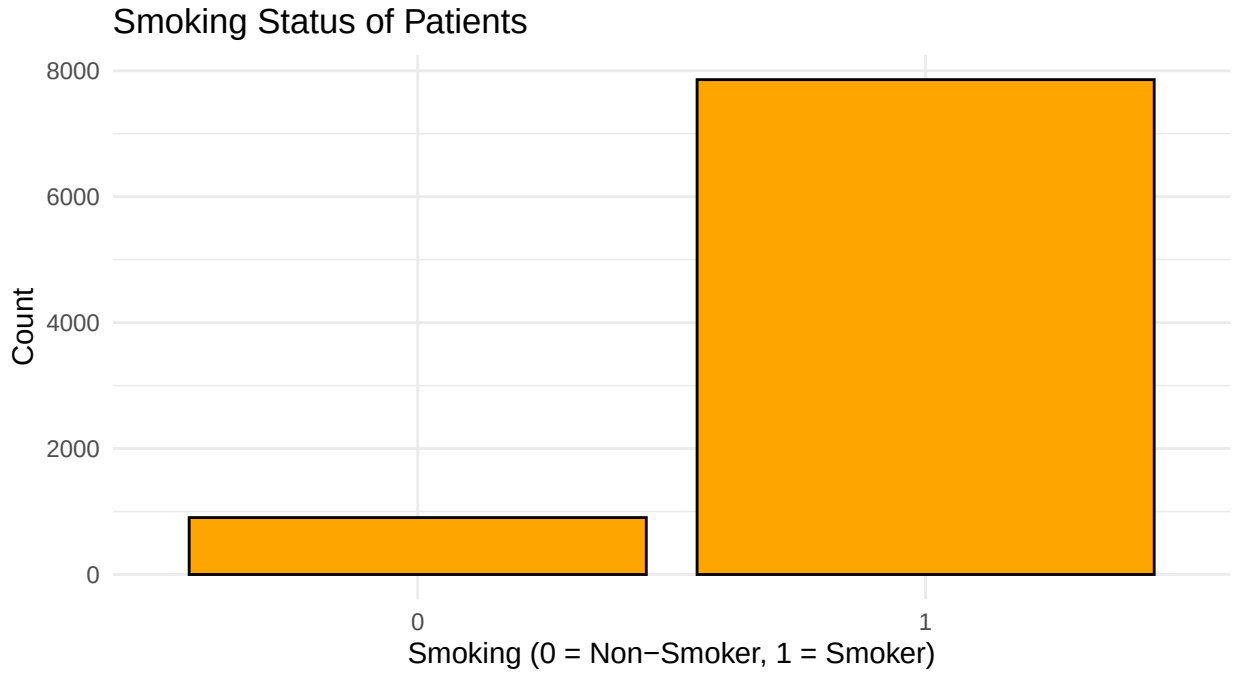
25 Dataset justification

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.

26 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.





27 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?

28 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.

29 Descriptive statistics and data visualizations

In this group of data, clinical and lifestyle measures show clear structure consistent with cardiovascular risk profiles. Age and BMI show right-skewed distributions with central tendencies in plausible clinical ranges (median age = 54 years; median BMI = 28.8 kg/m^2), while lipids such as cholesterol and triglycerides span wide ranges (see Appendix Table A1 for full summaries). Visual patterns display that individuals labeled at risk for heart attack tend to be older and have higher BMI relative to those not at risk (see Appendix Figures A1–A2). Proportional bar charts suggest higher risk prevalence among smokers and those with obesity compared with their counterparts (see Appendix Figures A4–A5), and risk proportions appear to differ by sex (see Appendix Figure A3).

The correlation heatmap (Appendix Figure A6) shows that body measurements and lipid levels are strongly related to each other, which makes sense since these factors are often linked in heart health. In contrast, variables tied to lifestyle and behavior only show weaker connections with vitals. Overall, these descriptive results give a good starting point for the statistical tests we plan to run (like chi-square for categorical variables and t-tests or ANOVA for continuous ones). They also suggest that using multivariable models could help us better understand how factors like age, body size, cholesterol, and lifestyle habits each contribute to heart attack risk.

30 Planned statistical methods

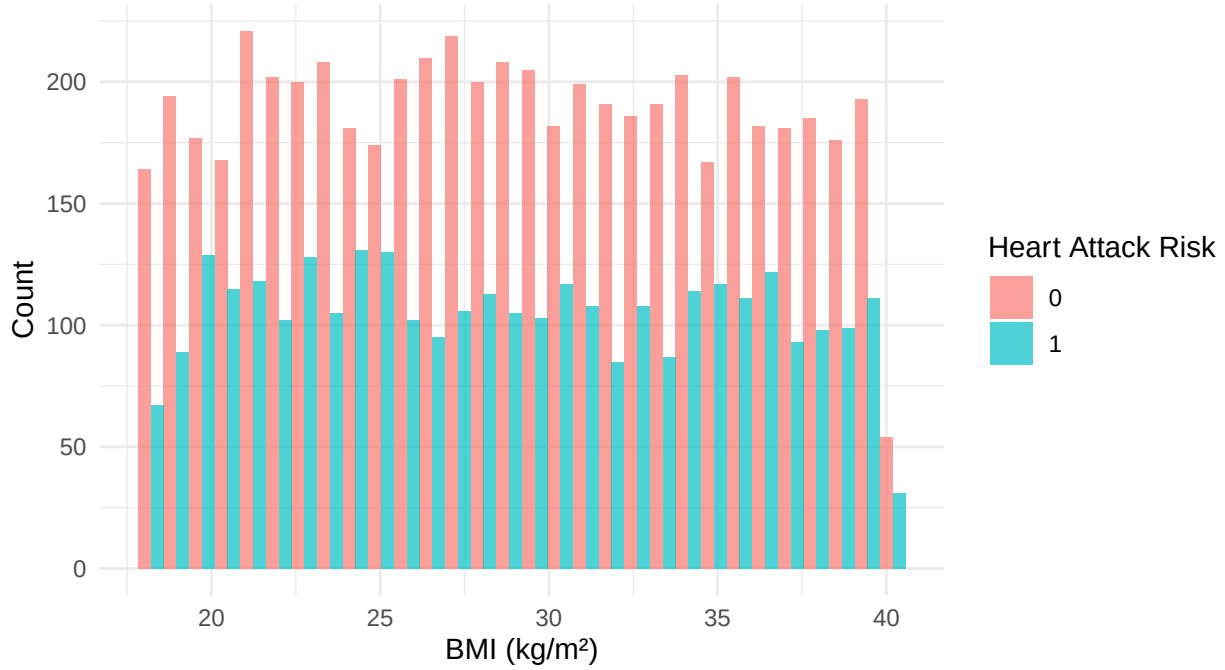
As the project progresses, I plan to use chi-square tests to assess associations between categorical factors (e.g., smoking, diabetes) and heart attack risk, and t-tests/ANOVA to compare continuous

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.

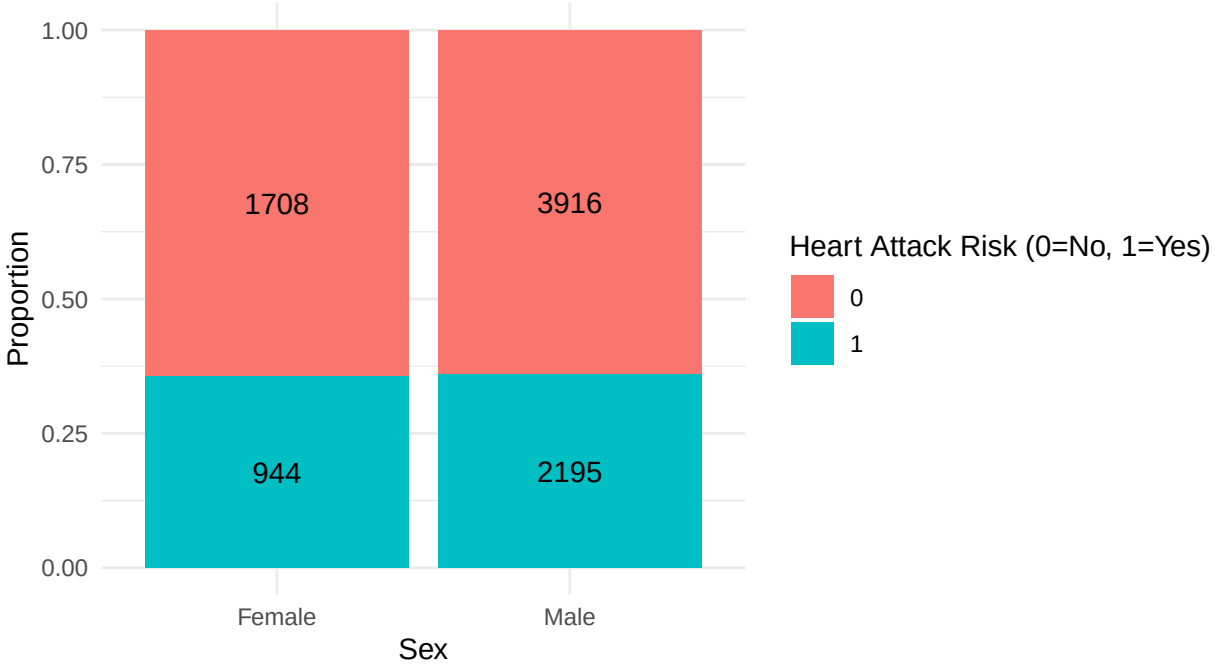
31 Limitations

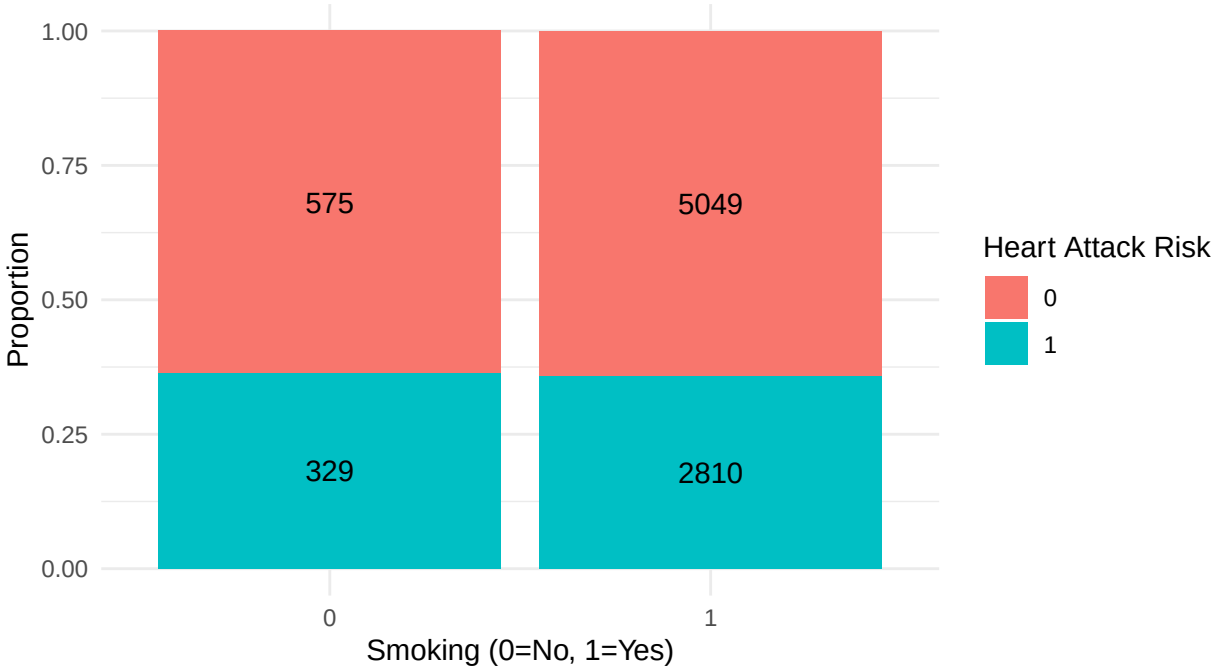
While descriptive statistics provide valuable insights into the dataset, they also have some limitations. Measures like mean and standard deviation can be influenced by outliers, which may distort the true central tendency and variability of the data. Categorical variables summarized with frequency counts 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.

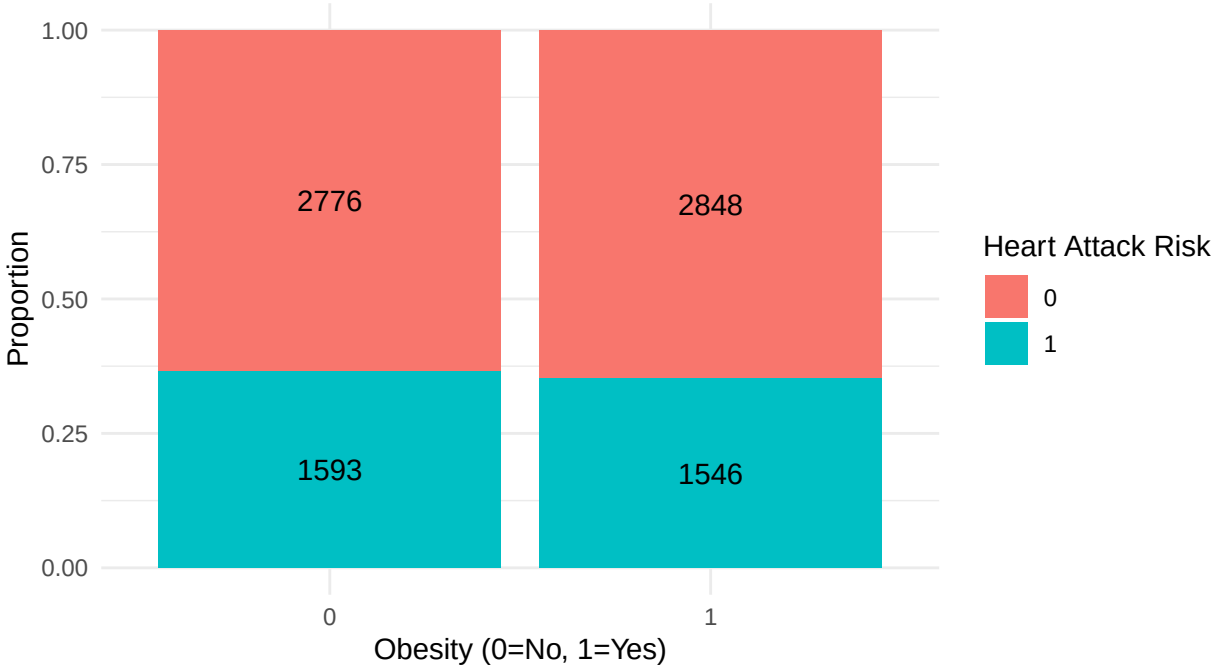
32 Appendix - Project Three

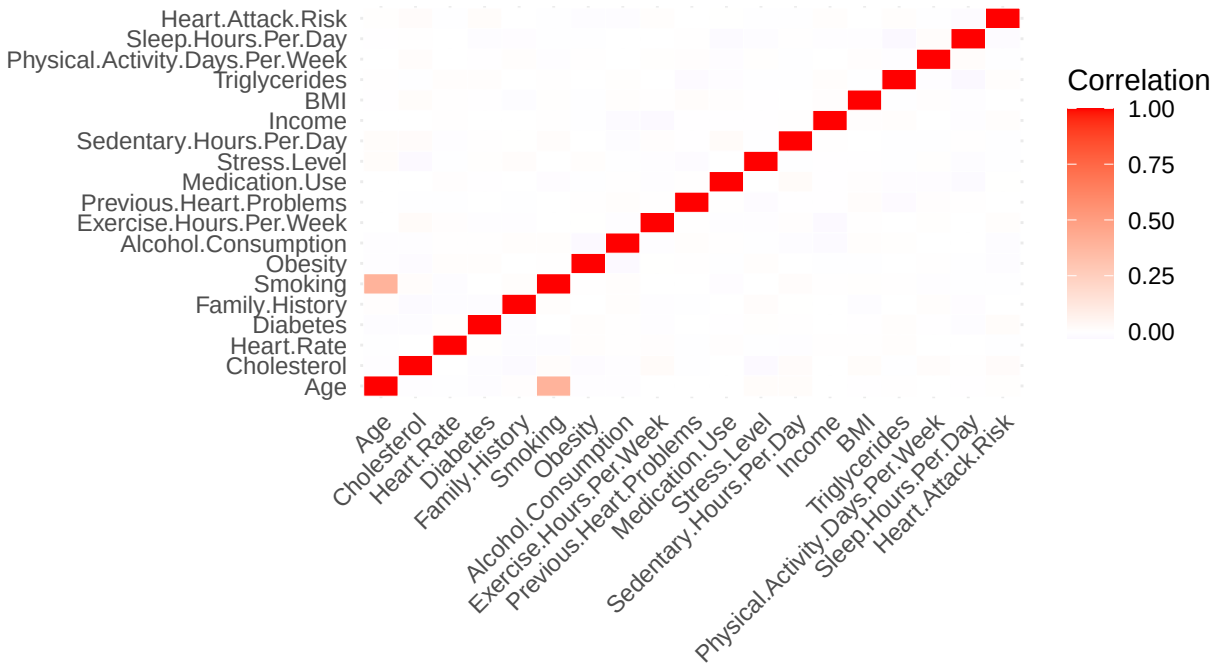



Count
200
100
0
25
50
75
Age (years)
Heart Attack Risk
0
1


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


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


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



33 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