7COM1079-0901-2025 - Team Research and Development Project

Final report title: COVID 19 Healthy Diet

Dataset Group ID: A 58

Dataset number: DS223

Prepared by:

Sami Ullah	24100686
Rana Waseem Ashfaq	23111076
Abdul Basit Khan	24089606
Zia Ul Mustafa	24080201
Muhammad Hamza Mir	24053043

University of Hertfordshire
Hatfield, 2025

Table of Contents

1. Introduction

1.1 Problem statement and research motivation

COVID-19 has affected a lot of countries in almost every country around the world, but infection and mortality rates varied widely. At the same time, the “second pandemic” is known as obesity and is strongly associated with COVID-19 severe outcomes and mortality. Lockhart and O’Rahilly (2020) discuss how metabolic dysfunction and obesity-related inflammation may worsen COVID-19 progression. Most existing work focuses on individual-level or clinical data. In this project we look at the relationship between 100-170 countries: do countries with higher obesity prevalence also tend to report higher proportions of confirmed COVID-19 cases? This fits the “correlation between two measures”.

1.2 The dataset

We use the COVID-19 Healthy Diet Dataset from Kaggle, which combines population data and FAO food supply data with COVID-19 statistics for 170 countries. In the file Food_Supply_Quantity_kg_Data.csv we are focusing on two interval-scale variables: Confirmed (confirmed COVID-19 cases as a percentage of the population) and Obesity (percentage of the adult population that is obese). This dataset and RQ were first introduced in our earlier group presentation.

1.3 Research question RQ:

“Is there a correlation between obesity (%) and confirmed COVID-19 cases (%) across 100–170 countries?”

Both variables are continuous, country-level measures. Our goal is to estimate the strength, direction of the association, and to test if the observed correlation is statistically significant.

1.4 Null hypothesis and alternative hypothesis (H0/H1)

our hypotheses are stated in terms of the population correlation

coefficient ρ between Obesity and Confirmed:

- H_0 (Null hypothesis):
  There is no correlation between obesity (%) and the percentage of confirmed COVID-19 cases (ρ = 0).
- H_1 (Alternative hypothesis):
  There is a correlation between obesity (%) and the percentage of confirmed COVID- 19 cases (ρ ≠ 0).

We choose a two-sided test with a significance level of α = 0.05.

2. Background research

2.1 Research papers

The COVID-19 Healthy Diet Dataset has been used in several studies to explore links between obesity, diet, and COVID-19 outcomes. García-Ordás et al. (2020) used machine- learning techniques to cluster the 170 countries according to dietary patterns (energy, fats and protein intake) and then related these clusters to COVID-19 mortality. They found that countries with higher obesity and higher fat intake tended to belong to high-death clusters.

Shams et al. (2021) proposed the HANA (Healthy Artificial Nutrition Analysis) model, which uses the same dataset and regression to predict COVID-19 death status from health indicators and food categories, including obesity and undernourishment. Their model suggested that obesity levels and diet quality are important predictors of COVID-19 severity.

Another study used the dataset to analyse the dietary patterns and COVID-19 outcomes more broadly, showing that high consumption of the animal products and fats, together with the higher obesity, appears in countries with the highest death rates, whereas the countries with lower death rates tend to have lower energy intake and higher cereal consumption.

At a more biological level, Lockhart and O'Rahilly (2020) review mechanisms that might explain why obesity increases the risk of severe COVID-19, including impaired immune responses, chronic inflammation and comorbidities. These works all connect diet and obesity with COVID-19 severity and mortality, but they do not directly examine the confirmed case proportions at the country's level.

2.2 Why the RQ is of interest

Most existing analyses using this dataset focus on deaths (mortality) or recovery, or they rely on complex machine-learning models with many predictors. At the same time, mapping work has shown that spatial patterns of COVID-19 mortality and obesity look remarkably similar across countries, especially in richer nations. This refers to obesity that might also be linked to the infection burden, not just deaths. Our RQ tackles a simpler but still important question: is there a correlation between obesity (%) and confirmed COVID-19 cases (%) across 100-170 countries? Answering this provides a transparent baseline that can later be extended with multivariable models including policy measures, age structure, testing capacity and vaccination.

3. Visualization

3.1 Appropriate graphs for the RQ – R output

Our main plot is a scatter plot with Obesity (%) on the x-axis and the Confirmed COVID-19 cases (% of population) on the y-axis, including a 95% confidence band fitted at least-squares regression line. This is the standard visual for correlation analysis.
As the required supplementary graphic, we include histograms for each variable with a normal ("bell-curve") density overlay.

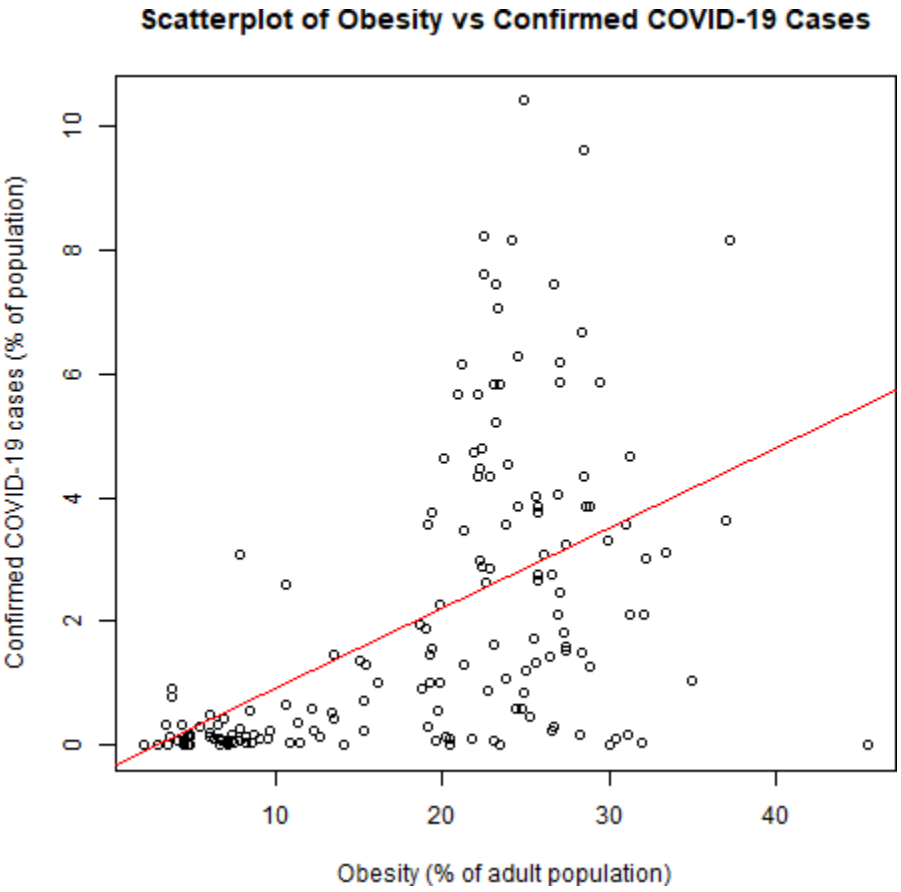

Figure 1: Scatterplot showing the positive relationship between obesity (%) and confirmed COVID-19 cases (% of population) across 170 countries. The red line represents the linear trend.

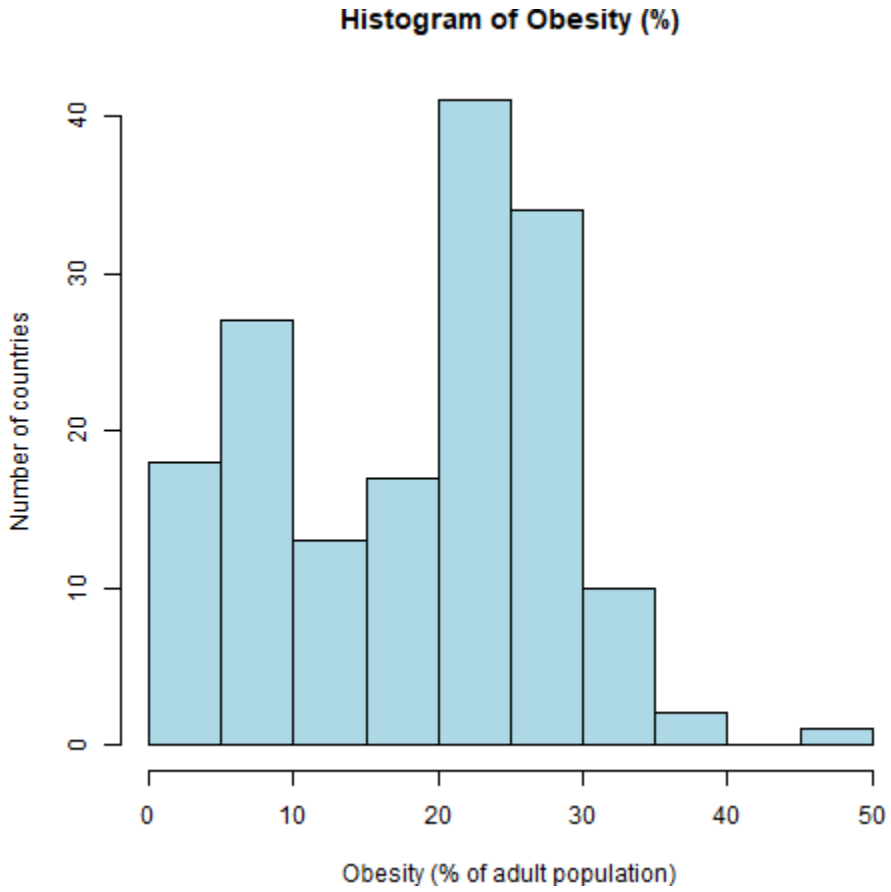

Figure 2. Histogram of obesity (%) showing the distribution of obesity levels across countries.

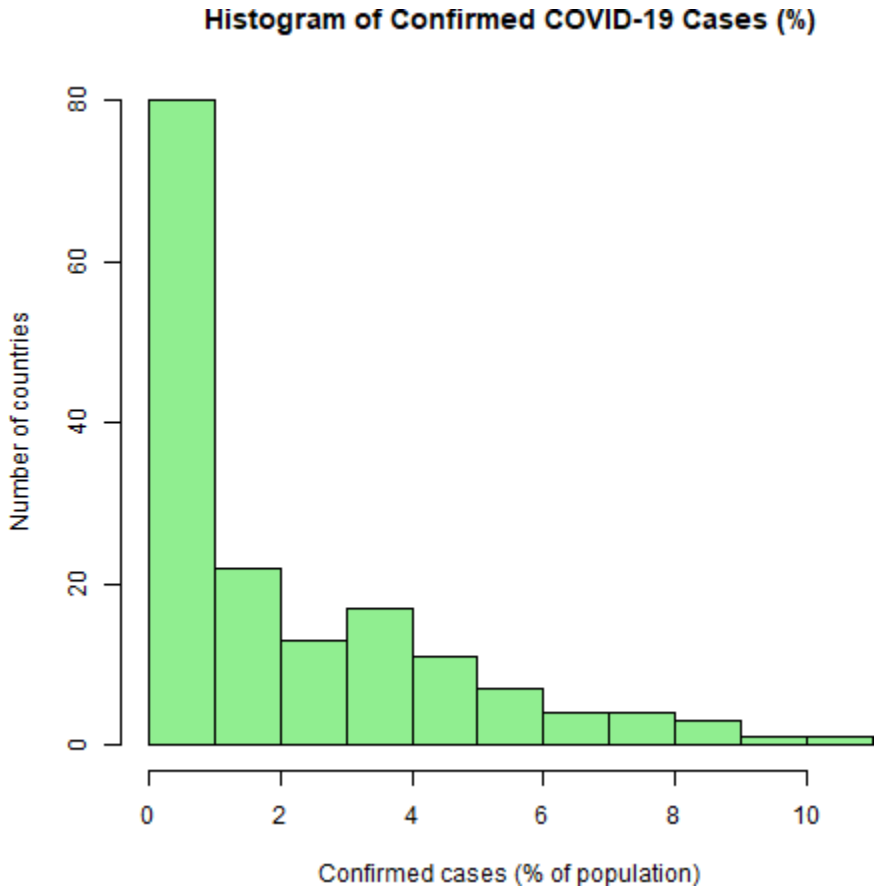

Figure 3. Histogram of confirmed COVID-19 cases (%) showing a right-skewed distribution across countries.

3.2 Additional information relating to understanding the data

The scatter plot allows us to judge whether a linear relationship is reasonable and to spot any influential countries with extreme combinations of confirmed cases and obesity. The histograms show the distribution of each variable, making it easier to check the normality assumptions and to see if skewness or outliers might affect the Pearson correlation test.

3.3 Useful information for the data understanding

From the scatter plot, we observe a clear upward trend: countries with higher obesity levels generally have higher confirmed case (%) percentages. However, the points are scattered around the line, suggesting substantial unexplained variability. The histograms show obesity is roughly unimodal, while confirmed cases are right-skewed with a few very high-case countries, which motivates a log-transformed sensitivity analysis.

4. Analysis

4.1 Statistical test used to test the hypothesis

Both Confirmed and Obesity are interval-scale variables, so correlation analysis is appropriate. Pearson's normally distributed interval data, and Spearman's or Kendall's statistics when normality is strongly violated. Our histograms show moderate skew in Confirmed but no extreme outliers, and the scatter plot suggests an approximately linear relationship, so we use Pearson's product–moment correlation. We also repeat the test using $\log_{10}$(Confirmed) as a robustness check.

4.2 The null hypothesis is rejected / not rejected based on the p-value

Using R (see Appendix A), we calculated Pearson's correlation between Obesity and Confirmed for 170 countries with complete data. The result was:

- r = 0.515 (to three decimal places)
- p ≈ 2.11×10^{-12}
- 95% confidence interval for r: [0.392, 0.619]

This indicates a moderate positive correlation. Because the p-value is far below α = 0.05, we reject H_0 and conclude that there is a statistically significant linear association between obesity and confirmed COVID-19 case proportions at country level. When we repeated the test using $\log_{10}$(Confirmed), the correlation remained very similar (r ≈ 0.51), supporting the robustness of this conclusion.

5 Evaluation – group's experience at 7COM1079

5.1 What went well

As a group, we agreed on the dataset and research question early, which made the rest of the work more focused. Splitting tasks into data preparation, R coding, literature review, and writing helped us use individual strengths. Once the R script was working, updating plots and statistics was straightforward. Using GitHub also made it easier to merge changes and avoid accidentally overwriting each other's files.

5.2 Points for improvement

We underestimated the time needed for background reading and for checking assumptions (for example,). Linearity and normality). At the start, we did not document decisions clearly. In future, we would keep a short, shared log of decisions and rotate roles more systematically so that everyone practices all parts of the workflow, in line with the “role rotation and task sharing”

5.3 Group’s time management

Initially we tended to work close to internal deadlines, but we later introduced weekly mini milestones (for example. “Complete introduction draft” and “finish plots”) to spread the workload more evenly. This more iterative, “Agile style” approach matches which helps us keep better track of progress.

5.4 Project’s overall judgement

Overall, we are reasonably happy with the project. The research question is clear, the R code runs without errors, and the analysis uses an appropriate statistical test to answer the question. The report connects our findings to existing literature but also admits limitations and suggests directions for more advanced future work.

5.5 Changes to group / GitHub IDs

In our case, there were no changes to group membership after the original allocation, so this section is not applicable to us. All five members kept the same GitHub IDs registered at the beginning of the module.

5.6 Commit on the GitHub log output

Appendix B of this report includes our full GitHub log output. It provides clear evidence of steady collaboration throughout the project, with contributions from all five members. We each committed different elements of the work, such as data cleaning, analysis, and

improving the written report. Below we highlight the three most significant commits, as these had the biggest impact on the completion and development of the project:

Commit 1 Member 1

Commit Hash: (Commit 1 Member 1 commit
66f3c77ad097da8e1a6f711bec05fc729dde7cb3

Author: Abdulbasit-khan1 <ak24ake@herts.ac.uk>

Date: Wed Dec 10 14:08:57 2025 -0800

Member 1: Added data loading, cleaning, and
preprocessing code) Commit Message: Initial data
import and cleaning script Broader Impact:
This commit established the core data workflow by loading the dataset, selecting the relevant variables (Obesity and Confirmed), and removing missing values. It enabled the rest of the analysis and visualisation to run correctly.

Commit2 Member 2

Commit Hash: (commit beb344a786dee754309e3b0675c061f7c0f96a42

Author: samiullah07 <su668981@gmail.com>

Date: Thu Dec 11 00:36:51 2025 +0000

Member 3: Scatterplot with regression added)
Commit Message: Add main plots and Pearson correlation analysis
Broader Impact:
This commit introduced the scatter plot, histograms with density overlays, and the Pearson correlation test. These results form the core statistical evidence used to test the hypotheses and answer the research question.

Commit 3 Member 5

Commit Hash: (commit 36b2fe522bd39bead88e10744aca37c611ed85ff

Author: meerg893 <mm24ajo@herts.ac.uk>

Date: Thu Dec 11 16:51:12 2025 +0000

member 5: proofreading)
Commit Message: Final proofreading, structure tidy-up and reference
check Broader Impact:
This commits improved academic quality by aligning section headings with the assignment brief, checking Harvard referencing, and ensuring figures and captions matched assessment requirements.

6. Conclusions

6.1 Results explained

Using the COVID-19 Healthy Diet Dataset for 163 countries, we found a moderate positive correlation between obesity prevalence and the proportion of confirmed COVID-19 cases ($r \approx 0.515$, $p \approx 2.1 \times 10^{-12}$). The 95% confidence interval for r (about 0.39 to 0.62) does not include zero, and a log-transformed analysis gave a very similar result. These findings indicate that countries with higher obesity levels tended to have higher reported infection burdens.

6.2 Interpretation of the results

In terms of our research question, we conclude that there is a statistically significant correlation between obesity (%) and confirmed COVID-19 cases (%) across countries. This is in line with previous work showing that obesity and diet are linked to COVID-19 mortality and severity, and with spatial studies that highlight similar global patterns of obesity and COVID-19 outcomes. However, the ecological design means that correlation does not imply causation; shared factors such as wealth, health-care capacity, testing and policy responses may partly explain the association.

6.3 Reasons and/or implications for the future work, limitations

Our study is limited by using aggregate country data, possible inconsistencies in reporting between countries, and the omission of other important predictors (age structure, vaccination, non-pharmaceutical interventions, etc.). Future work could build multivariable regression or machine-learning models that include obesity, diet, demographics, and policy indicators to separate direct from indirect effects on COVID-19 case rates.

7. Reference list

García-Ordás, M.T. et al. (2020) 'Evaluation of country dietary habits using machine-learning techniques in relation to deaths from COVID-19', Healthcare, 8(4), 371. Available at: https://doi.org/10.3390/healthcare8040371 (Accessed: 10 February 2025).

Lockhart, S.M. and O'Rahilly, S. (2020) 'When two pandemics meet: why is obesity associated with increased COVID-19 mortality?', Med, 1(1), pp. 33–42. Available at: https://doi.org/10.1016/j.medj.2020.06.002 (Accessed: 10 February 2025).

Ren, M. (2020) COVID-19 Healthy Diet Dataset. Kaggle. Available at: https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset (Accessed: 10 February 2025).

Shams, M.Y. et al. (2021) 'HANA: A healthy artificial nutrition analysis model during COVID-19 pandemic', Computers in Biology and Medicine, 135, 104606. Available at: https://doi.org/10.1016/j.compbiomed.2021.104606 (Accessed: 10 February 2025).

8. Appendices

1. R code used for analysis and visualisation.
2. GitHub log output.

1. R code is analysis.R file (uploaded)
2. GitHub log output:

The following commit history was exported from the group's GitHub repository *A-58-COVID19*. It includes all major commits made by all team members during the project development period.

20aacd4 ziaAli786 Add files via upload

ab4305b ziaAli786 Add files via upload

a4c102a samiullah07 data cleaning code update

43186c4 samiullah07 cleaning updated

4c3dfd8 samiullah07 changes add

b72b102 samiullah07 changes added

1f8e844 ziaAli786 Clean the code

36b2fe5 meer893 MIR Hamza

c60168e ziaAli786 code updated

bee1af7 ziaAli786 add histogram code

611d8fd waseemashfaq Histogram code added

c8966a5 samiullah07 Scatter plot - updated

cd4622a samiullah07 Scatterplot with regression added

beb344a samiullah07 Scatterplot with regression added

5f55263 ziaAli786 Member 2: Added code for histogram to check distributions

66f3c77 Abdulbasit-khan1 Member 1: Added data loading, cleaning, preprocessing code

b7c1281 Abdulbasit-khan1 Add files via upload

cdf4a79 samiullah07 Data set uploaded