Data Wrangling – Project Report

Meat Consumption and Mortality Rate

Group 3

Joana Petkova	**Mahbod Tajdini**	**Selman Gül**	**Takehiro Tsurumi**
jpe206	mta266	sgu207	tti440
2757809	2756234	2756789	2721092

1. Introduction

Numerous studies have examined the connection between meat consumption and human health with an emphasis on its impact on mortality. Vital nutrients, including proteins, vitamins, and minerals, are found in meat; nevertheless, high intake of some forms of meat—particularly red and processed meats—has been associated with negative health effects (González et al., 2020; Larsson & Orsini, 2013). These relationships have been examined by epidemiological research, which highlights the need to understand how different types of meat influence mortality rates.

The European Prospective Investigation into Cancer and Nutrition (EPIC) study examined 448,568 individuals aged 35 to 69 to determine if meat consumption was linked to premature mortality. The findings revealed that high red meat intake was associated with increased all-cause mortality, with an even stronger correlation for processed meat. According to the study, lowering the daily intake of processed meat to less than 20g could prevent 3.3% of deaths. Notably, poultry intake and all-cause mortality did not significantly correlate (Rohrmann et al., 2013).

Similarly, a prospective study that included more than half a million people between the ages of 50 and 71 looked into the connection between meat consumption and mortality. The results showed that those with the highest red meat intake were at a higher risk of death. Processed meat consumption was also linked to higher mortality risk. On the other hand, more white meat intake was associated with an overall lower chance of death (Sinha et al., 2009).

Additional research has investigated the effects of red meat on cardiovascular disease risk, particularly in women (Bovalino et al., 2016), as well as its role in aging (Kouvari et al., 2016). Furthermore, frequent consumption of red and processed meat has been linked to an increased risk of chronic illnesses, which highlights the need for a balanced diet and moderate consumption (González et al., 2020).

Although these studies offer valuable insight into the health effects of eating red and processed meat, less is known about how other meats—such as poultry, fish and seafood—impact fatalities. Moreover, regional studies are required due to differences in meat consumption behaviors. Thus, this project aims to investigate the following research question:

How does the consumption of various types of meat (poultry, beef, fish and seafood, pork, sheep and goat, other types) relate to mortality rates across European countries from 2009 to 2019?

Additionally, this work addresses the following sub-questions:

1. *Is there a correlation between the consumption of different types of meat and mortality rate from cardiovascular disease in European countries during the same period?*
2. *Which type of meat shows the strongest correlation with overall mortality rate and mortality rate from cardiovascular disease?*

To answer these questions, various data wrangling techniques were applied to three datasets, enabling the analysis of regional differences in these correlations and offering insights into the impact of meat intake on overall mortality rate and mortality rate from cardiovascular diseases. Understanding these relationships is crucial for developing targeted dietary recommendations and health campaigns aimed at reducing diet-related mortality risks.

2. Data Sources

Three separate datasets were used for this project. The first dataset contained data about the consumption of different types of meat (poultry, beef, fish and seafood, pork, sheep and goat, other types) in nearly all countries worldwide from 1961 to 2021. The dataset was downloaded in CSV format from the website Kaggle on January 21, 2025, and originally sourced from the Food and Agriculture Organization of the United Nations (Scibearia, 2024). The second dataset included global mortality data from 2009 to 2021, covering overall mortality rates as well as gender-specific rates, with region labels for each country. It was downloaded from the website of the World Health Organization's data platform on January 26, 2025, also in CSV format (World Health Organisation, *Causes of death explorer* 2025). The third dataset consisted of the mortality rates due to cardiovascular diseases in almost all countries worldwide from 1961 to 2021. It was downloaded from the website Our World in Data on January 21, 2025, but the data was originally

provided by the World Health Organization (Dattani et al., 2023). The last access date of each webpage was the day on which the corresponding dataset was downloaded.

3. Scope

For the purpose of this work, countries are classified as European based on the categorization of the United Nations Statistical Division (United Nations, 2025). Territories that are not sovereign states are not considered. The datasets utilized contain data about all European countries except Andorra, Lichtenstein, Monaco, and San Marino.

The years 2020 and 2021 are not examined in this project due to the COVID-19 pandemic, which was ongoing at the time. The reason is that the pandemic caused an increase in mortality rates, which would introduce an external factor that could distort the analysis and correlations in this study.

The mortality rates used in this work are per 100 000 population and standardized according to the age distribution of the corresponding country's population. This standardization ensures comparability across countries with different age structures, as countries with similar age-specific mortality rates may show differing overall mortality rates without this adjustment (World Health Organization, *Age-standardized mortality rate (per 100 000 population)* 2025).

4. Data Wrangling Methods

Data wrangling is a key step in preparing raw data for analysis. In this project, a systematic data preprocessing workflow was applied to ensure that the datasets are properly formatted and contain the necessary information for meaningful analysis.

4.1 Data Acquisition

The three datasets used in this work were found through targeted searches for data from established sources of statistical information, including specialized agencies of the United Nations and the World Health Organization. During the data acquisition process, unnecessary information was filtered out, particularly when downloading the dataset with overall mortality rates from the World Health Organization's data platform. All datasets were acquired in the form of CSV files and parsed using the read_csv() method available in the Pandas library to transform them into structured tabular format for analysis. To ensure consistency and avoid potential issues, proper character encoding was applied during the parsing process. In addition, the unnecessary metadata available at the start of the overall mortality rate dataset was omitted to ensure correct conversion to tabular format.

4.2 Data Cleaning and Merging

We began the data cleaning process by removing unnecessary columns and renaming those with inconsistent or unclear names to enhance clarity for later processing. The null values present across all datasets were examined and found to be irrelevant, as they would be eliminated when filtering the datasets to focus on the time period and countries relevant to the investigated questions. However, upon attempting to merge the overall mortality rate and cardiovascular disease mortality rate datasets with the meat consumption dataset, it became evident that data for some years for certain countries was missing due to the emergence of null values.

The missing values were imputed using quadratic regression models, as opposed to simple mean imputation. This approach was chosen because the mortality rates in both datasets showed a decreasing trend over the analyzed time period, and regression provided more accurate estimates that corresponded to the trend. Countries with excessive amounts of missing data—specifically Albania, Bosnia and Herzegovina, and Montenegro—were excluded from the analysis at this stage. The missing rows were added using the Pandas concat() function, and all dataframes were merged with the Pandas merge() method, ensuring that data instances aligned correctly based on shared country identifiers and years while preserving the integrity of the original dataframes.

4.3 Data Aggregation and Visualization

Once the datasets were cleaned and merged, additional transformations were applied to enhance their usability. Versions of the meat consumption, mortality, and cardiovascular disease mortality dataframes aggregated by country and year were made using the Pandas groupby() function. The mean values across countries or years were then calculated to generate various visualizations. Reshaping techniques, such as the melt() function, were employed to convert between wide and long formats to make it easier to conduct different types of analysis.

The libraries Seaborn and Matplotlib were used to create a range of visualizations. Line graphs were generated to examine trends in meat consumption, overall mortality rates, and cardiovascular disease mortality rates over time. Stacked bar charts were utilized to depict the distribution of meat consumption and gender-specific mortality rates across countries. Pie charts were made to illustrate the gender distribution of overall mortality rates in Europe for the examined years. Horizontal bar charts were created to showcase the top countries for consumption of different meat types and the two mortality rates. Scatter plots were utilized to visualize the regression used for imputing missing values, as well as to explore potential relationships between meat consumption and mortality rates, both overall and from cardiovascular diseases. Additionally, scatter plots helped depict the clustering performed in later stages of the project.

Finally, the Folium library was used to create interactive maps that showcase geographical variations in overall mortality and cardiovascular disease mortality rates across European countries. This mapping technique allows for a spatial perspective on the data, helping to identify regional patterns and potential clusters in dietary habits and health outcomes. Many of these visualizations provided insight into underlying patterns in the data and, thus, guided our selection of appropriate statistical analyses.

4.4 Data Analysis

Both overall mortality and cardiovascular mortality rates showed a decreasing trend across all countries, which could interfere with the evaluation of the research questions. Since the influence of this trend falls outside the scope of this project, a detrending method was applied, where the residuals from the linear regression line were extracted from the entire mortality rate columns. These detrended values were used in the subsequent analysis to ensure unbiased statistical results.

The normality of the detrended mortality rates was checked both quantitatively, using the Shapiro-Wilk test, and visually, via a QQ-plot. Based on these results, an appropriate correlation test was selected to explore the relationship between these rates and the consumption of different types of meat. Since the data did not follow a normal distribution, Spearman's rank correlation was applied to assess monotonic relationships.

To further investigate whether specific meat consumption patterns influence the detrended overall and cardiovascular mortality rates, countries were clustered into three groups based on their consumption pattern using K-Means. A suitable statistical test was chosen based on the number of clusters, the normality of rates within each cluster (assessed with the Shapiro-Wilk test), and the homogeneity of variance within clusters (evaluated using the Levene test). While the rates within the clusters were homogeneous, the normality assumption did not hold, so the Mann-Whitney U test was applied to test for statistical significance between the mean overall and cardiovascular disease mortality rates of the clusters.

5. Findings

Considering that the ***detrended mortality rate (DMR)*** and ***detrended cardiovascular mortality rate (DCMR)*** did not follow a normal distribution, the Spearman correlation test was run. The results, displayed in Tables 1 and 2, indicate an overall trend where higher meat consumption is linked to lower ***DMR*** and ***DCMR***. In particular, "Beef", "Sheep and Goat", "Other Meats", and "Fish and Seafood" show greater statistical significance of the negative correlation ($p < 0.01$), with "Beef" exhibiting the strongest inverse relationship with both ***DMR*** and ***DCMR***. In contrast, "Poultry" showed no correlation, while "Pork" demonstrated a weaker negative correlation. These findings suggest that there is a relationship between specific meat-type consumption and mortality rates, though not all meat types show a strong statistical correlation.

Table 1: Spearman Correlation Table - Detrended Mortality Rate Across Different Meat Types (***<0.01, **<0.05, *<0.1)

	Poultry	Beef	Sheep and Goat	Pork	Other Meats	Fish and Seafood
Corr	0.05	-0.68	-0.48	-0.12	-0.58	-0.63
P-value	$3.09\text{x}10^{-1}$	$1.45\text{x}10^{-54}$	$1.55\text{x}10^{-24}$	$1.34\text{x}10^{-2}$	$1.68\text{x}10^{-37}$	$2.22\text{x}10^{-45}$
Sig.	No	***	***	**	***	***

Table 2: Spearman Correlation Table - Detrended Mortality Rate from Cardiovascular Diseases Across Different Meat Types (***<0.01, **<0.05, *<0.1)

	Poultry	Beef	Sheep and Goat	Pork	Other Meats	Fish and Seafood
Corr	-0.04	-0.70	-0.47	-0.08	-0.54	-0.65
P-value	$4.25\text{x}10^{-1}$	$2.62\text{x}10^{-60}$	$5.84\text{x}10^{-23}$	$9.25\text{x}10^{-2}$	$5.18\text{x}10^{-31}$	$7.92\text{x}10^{-50}$
Sig.	No	***	***	*	***	***

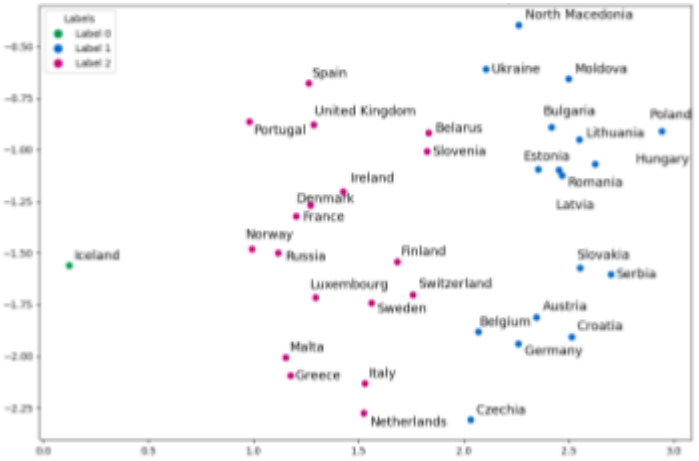

Fig 1: Clustering of Countries with K-Means (random_state=42) based on Meat Consumption Patterns

To further examine whether meat consumption patterns affect ***DMR*** and ***DCMR***, countries were clustered into three groups based on their meat consumption profiles, as shown in Fig.1. Notably, Iceland formed a distinct cluster due to its exceptionally high consumption of "Fish and Seafood" and "Sheep and Goat". The presence of a single country in this cluster made conducting a statistical test with it unfeasible; thus, Iceland was excluded from the subsequent analysis.

Among the remaining clusters, Group 1 (Label 1) has a higher ***DMR*** and ***DCMR*** alongside slightly greater "Pork" consumption, while Group 2 (Label 2) has lower ***DMR*** and ***DCMR*** and higher consumption of "Beef" and "Fish and Seafood". Normality and variance homogeneity tests confirmed that neither group met the assumptions of normality or equal variance in terms of both ***DMR*** and ***DMCR***. Specifically, for ***DMR***, the Shapiro-Wilk test yielded p-values of 0.343 for Group 1 and $3.531\text{x}10^{-6}$ for Group 2, and the Levene test produced a p-value of 0.342. For ***DMCR***, the Shapiro-Wilk test returned p-values of 0.869 and $2.964\text{x}10^{-6}$ for Groups 1 and 2, respectively, with a Levene test p-value of 0.116. Given these results, the Mann-Whitney U test was used to find if ***DMR*** and ***DCMR*** were significantly higher in Group 1 than in Group 2. The test results ($U = 263$, $p = 0.0002$ for ***DMR***; $U = 264$, $p = 0.0001$ for ***DCMR***) confirmed that Group 1 has significantly larger ***DMR*** and ***DCMR*** than Group 2.

These findings align with the earlier correlation analysis, where "Beef" and "Fish and Seafood" showed the strongest negative correlations with mortality rates. The statistically significant differences in ***DMR*** and ***DCMR*** between the two

clusters suggest that higher consumption of "Beef" and "Fish and Seafood" is linked to lower mortality rates.

6. Conclusion and Future Work

This data wrangling project demonstrated that there is a relationship between the consumption of certain meat types and overall and cardiovascular disease mortality rates in European countries across 2009 to 2019. More specifically, a strong negative correlation was observed for beef, sheep and goat, fish and seafood, and other meats, with beef consumption showing the strongest correlation. Clustering analysis further revealed that European countries with higher consumption of beef, fish, and seafood tend to have lower overall and cardiovascular mortality rates. These findings suggest potential health benefits linked to specific meat types, aligning with evidence that fish and lean meats may support cardiovascular health (González et al., 2020; Sinha et al., 2009). However, the observed inverse relationship for red meats like beef contrasts with prior studies (Rohrmann et al., 2013; Larsson & Orsini, 2013), possibly due to regional dietary practices or data aggregation that conflated processed and unprocessed meat categories.

This work has several limitations. The use of nationally aggregated data may mask individual consumption patterns and introduce measurement errors. Additionally, unmeasured confounders, such as socioeconomic factors, healthcare access, and broader lifestyle habits, could bias the results. The exclusion of processed meat as a distinct category limits comparability with studies emphasizing its risks. Furthermore, omitting 2020–2021 data due to the COVID-19 pandemic restricts insights into more recent meat consumption-mortality trends.

Future research should prioritize individual-level longitudinal studies to clarify causal relationships and control for confounders. Differentiating between processed and unprocessed meats, as well as cooking methods could reveal the mechanisms behind the correlations observed in this study. Integrating genetic data may also help explain regional trends. Expanding the scope to post-2019 data and non-European regions would test the generalizability of the findings, while exploring interactions between meat and other dietary components could contextualize the results. Addressing these gaps would strengthen the evidence base for dietary guidelines, helping to balance public health priorities with regional and nutritional diversity. This would allow researchers to improve strategies for reducing mortality risks, while recognizing the complex role of meat in global diets.

References

Bovalino, S., Charleson, G. and Szoeke, C. (2016) 'The impact of red and processed meat consumption on cardiovascular disease risk in women', *Nutrition*, 32(3), pp. 349–354. doi:10.1016/j.nut.2015.09.015.

Dattani, S. *et al.* (2023) *Cardiovascular diseases*, *Our World in Data*. Available at: https://ourworldindata.org/cardiovascular-diseases?insight=death-rates-from-cardiovascular-diseases-have-declined-in-many-countries#key-insights (Accessed: 21 January 2025).

González, N. *et al.* (2020) 'Meat consumption: Which are the current global risks? A review of recent (2010–2020) evidences', *Food Research International*, 137, p. 109341. doi:10.1016/j.foodres.2020.109341.

Kouvari, M., Tyrovolas, S. and Panagiotakos, D.B. (2016) 'Red meat consumption and healthy ageing: a review', *Maturitas*, 84, pp. 17–24. doi:10.1016/j.maturitas.2015.11.006.

Larsson, S.C. and Orsini, N. (2013) 'Red meat and processed meat consumption and all-cause mortality: a meta-analysis', *American Journal of Epidemiology*, 179(3), pp. 282–289. doi:10.1093/aje/kwt261.

Rohrmann, S. *et al.* (2013) 'Meat consumption and mortality - results from the European Prospective Investigation into Cancer and Nutrition', *BMC Medicine*, 11(1). doi:10.1186/1741-7015-11-63.

Scibearia (2024) *Meat consumption per capita*, *Kaggle*. Available at: https://www.kaggle.com/datasets/scibearia/meat-consumption-per-capita/data (Accessed: 21 January 2025).

Sinha, R. *et al.* (2009) 'Meat intake and mortality: a prospective study of over half a million people', *Archives of Internal Medicine*, 169(6), pp. 562–571. doi:10.1001/archinternmed.2009.6.

United Nations (2025) *UN Statistical Division - Methodology*, *Geographic Regions*. Available at: https://unstats.un.org/unsd/methodology/m49/ (Accessed: 30 January 2025).

World Health Organisation (2025) *Causes of death explorer*, *World Health Organization: Data Platform*. Available at: https://platform.who.int/mortality/themes/theme-details/MDB/all-causes (Accessed: 26 January 2025).

World Health Organization (2025) *Age-standardized mortality rate (per 100 000 population)*, *World Health Organization*. Available at: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/78 (Accessed: 30 January 2025).