Is Food Making You Sick?

Exploring the relationship between nutrition and health

https://ykmsdv.github.io/thesis/

Yoana Kosturska
MS Data Visualization, Parsons The New School For Design

Submitted in partial fulfillment of the requirements for the degree of Master Science in Data Visualization at Parsons School of Design.

Faculty:
Daniel Sauter

May 20222

Abstract

Is Food Making You Sick? is exploring the relationship between diet and health in 184 countries through the analysis and visualization of extensive data on nutritional factors and diseases. The project aims to study and visualize these relationships by juxtaposing 47 nutrients and 84 health conditions, allowing users to explore and analyze them.

Is Food Making You Sick? will answer questions such as do countries that have a high fat diet, also struggle with heart disease? Are countries in which people are consuming high amounts of sugar also overburdened with diabetes? and many others. By exploring the patterns between food consumption and health outcomes, users could better inform their dietary choices and be empowered by high quality, research data rather than speculative data that often float around the Internet. Exposing the data visually could also spark discussion amongst researchers about less researched correlations, setting the stage for testing hypotheses and performing additional studies.

Introduction

"Let food be thy medicine and medicine be thy food."
Hippocrates

"No disease that can be treated by diet should be treated with any other means."
Maimonides

For the past two years, the world was overtaken by an unprecedented global pandemic that changed our behaviors, economies, and uprooted our lives. From the start, scientists and

healthcare workers were trying to answer the question: how dangerous is this new virus? Soon, theories were developed as we were seeing that patient's outcomes were vastly different – some people ended up in hospitals on ventilators while others were completely asymptomatic.

What made these outcomes different was attributed first to age, then to obesity and to preexisting conditions such as diabetes and asthma, and even behaviors such as smoking. While this information helped medical workers better judge and predict the likelihood of someone developing a serious case of Covid-19 and consequently distributed vaccines according to individual risk factors, for people there was nothing much we could do. We were either in the risk categories or not.

Covid-19 is just the latest example of how preexisting conditions affect people. According to WHO, the top 10 causes of death[1] fact sheet, "The world's biggest killer is ischaemic heart disease, responsible for 16% of the world's total deaths. [...] Diabetes has entered the top 10 causes of death, following a significant percentage increase of 70% since 2000. Diabetes is also responsible for the largest rise in male deaths among the top 10, with an 80% increase since 2000."

Diabetes, cardio-vascular disease, and other comorbidities are not only directly affecting the quality of life of patients, but as seen throughout the COVID-19 pandemic, are exponentially increasing the risk of negative outcomes from other diseases.[2]

What diabetes and cardio-vascular disease have in common with COVID-19 is their relationship to diet. It is well established that a high carb, high sugar diet increases the risk of diabetes.

[1] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
[2] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Coupled with chronic overeating and genetic predisposition, obesity is often a precursor or a comorbidity of diabetes. Then, cardio-vascular disease is heavily influenced by BMI, cholesterol, and level of activity. While, once faced with a disease or a virus, we are powerless to change our risk factors, we can proactively make choices in advance that better prepare us for these challenges.

Is Food Making You Sick? is studying the relationship between food consumption, portion size, and several health indicators in 184 countries, in order to shed light on the direct impact nutrition has on health. It tries to answer questions such as *Is the food we are eating making us sick? Is traditional food in countries related to the prevalence of certain diseases?* The purpose of this study is to explore, study, and visualize the relationship between the food we consume and our health by juxtaposing nutrition and health metrics. Visualizing the relationships can allow users to easily make assessments and consider whether they would want to augment their eating habits accordingly. Exposing the data visually could also spark discussion amongst researchers about less researched correlations, setting the stage for testing hypotheses and performing studies.

Background

Medicine clusters health conditions into two main groups, communicable and noncommunicable diseases. Communicable diseases are conditions that are infectious, bacterial or viral in nature, such as tuberculosis and COVID-19. Noncommunicable diseases are not contagious and include conditions such as stroke and diabetes.

The percentage of people living with diabetes in the United States increased from 0.93% in 1958 to 7.40% in 2015. In 2017, the death rate from cancer increased by 17% compared to 1990, and the share of population with breast and colon and rectum cancers were the highest, with 0.24% and 0.15% of the world population respectively[3]. This equates to an estimated 27 million people living with just those two types of cancer. Scientific literature often points to poor diet as one of the risk factors for some types of cancer, especially colon and rectum, pancreatic, or breast cancer in women in menopause.[4]

While we observe this steep increase in recent history, there is nothing new about the increase of noncommunicable diseases worldwide. The top leading causes of death have changed drastically in the past century – in the beginning of the 1900s, diseases associated with infections or bacteria, such as pneumonia, bronchitis, senility, nephritis, tuberculosis, and diarrheal diseases were the leading causes of death.

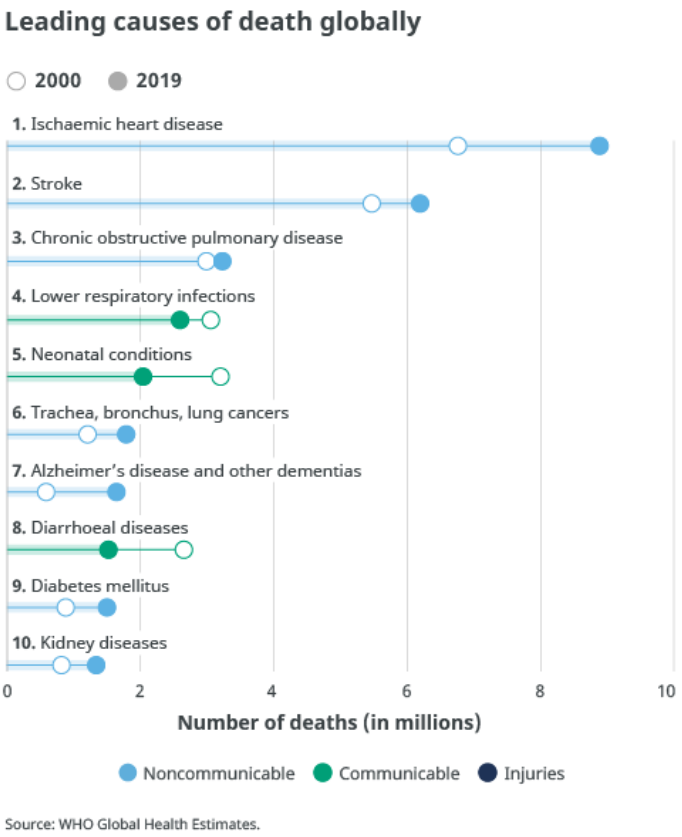
Then throughout the 20th century, as science and medicine progressed, the leading causes of death in the United States and the rest of the world slowly changed. By the 1950s noncommunicable diseases, such as ischaemic heart disease, stroke, chronic obstructive pulmonary disease and Alzheimer's disease and other dementias, were taking over and dominating the top ten leading causes of mortality across the world.[5]

---

[3] https://ourworldindata.org/cancer

[4] https://www.cancer.org/healthy/eat-healthy-get-active/acs-guidelines-nutrition-physical-activity-cancer-prevention/diet-and-activity.html

[5] https://www.cdc.gov/nchs/data/dvs/lead1900_98.pdf

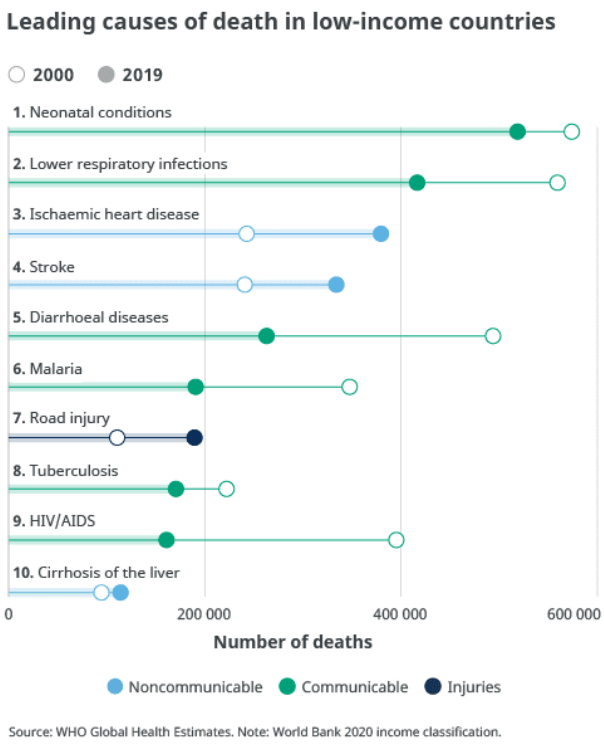


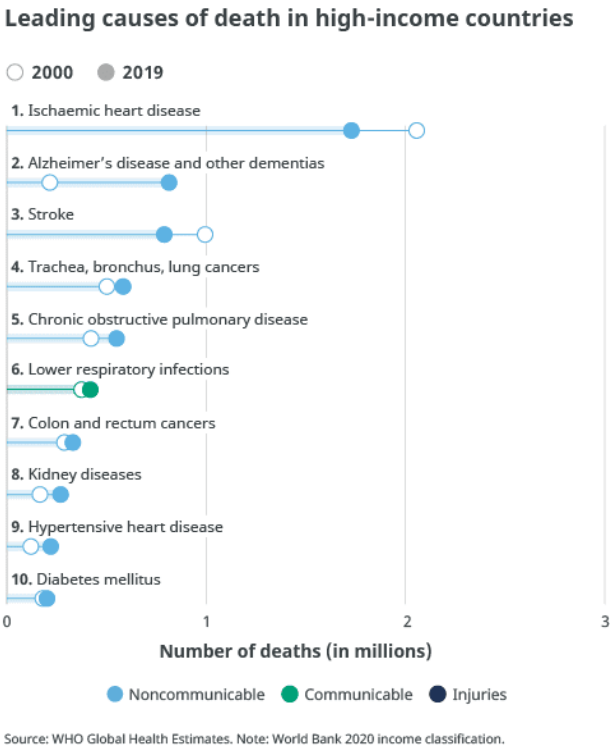
Nowadays, noncommunicable diseases cause more and more deaths every year, and some of the main risk factors associated with them are directly related to poor diet and lack of physical activity. In 2019, seven out of the ten leading causes of death were noncommunicable diseases. Just those seven causes accounted for 44% of all deaths, and combined with all noncommunicable diseases not part of the top ten causes, accounted for 74% of all deaths in the world.[6]

When one accounts for factors such as GDP and economic development, we observe that the higher the income of a country, the more noncommunicable diseases prevail. In lower income

[6] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

countries, communicable diseases are far more prevalent than in higher income countries, with six out of the top ten leading causes of death being communicable diseases, such as malaria, tuberculosis, and HIV/AIDS. Following the global trend we observe that deaths from communicable diseases decrease in all income levels, while deaths from noncommunicable diseases increase.





We all intuitively know this disparity exists and some of the reasons behind it - poor sanitation and maintenance in lower income areas, lack of education about transmission of diseases such as HIV/AIDS, lack of access to healthcare and needed medications. We also know why communicable diseases are stepping down from being the top causes of death globally, and the

statistics support this intuition by showing us just how quickly noncommunicable diseases are taking prevalence.

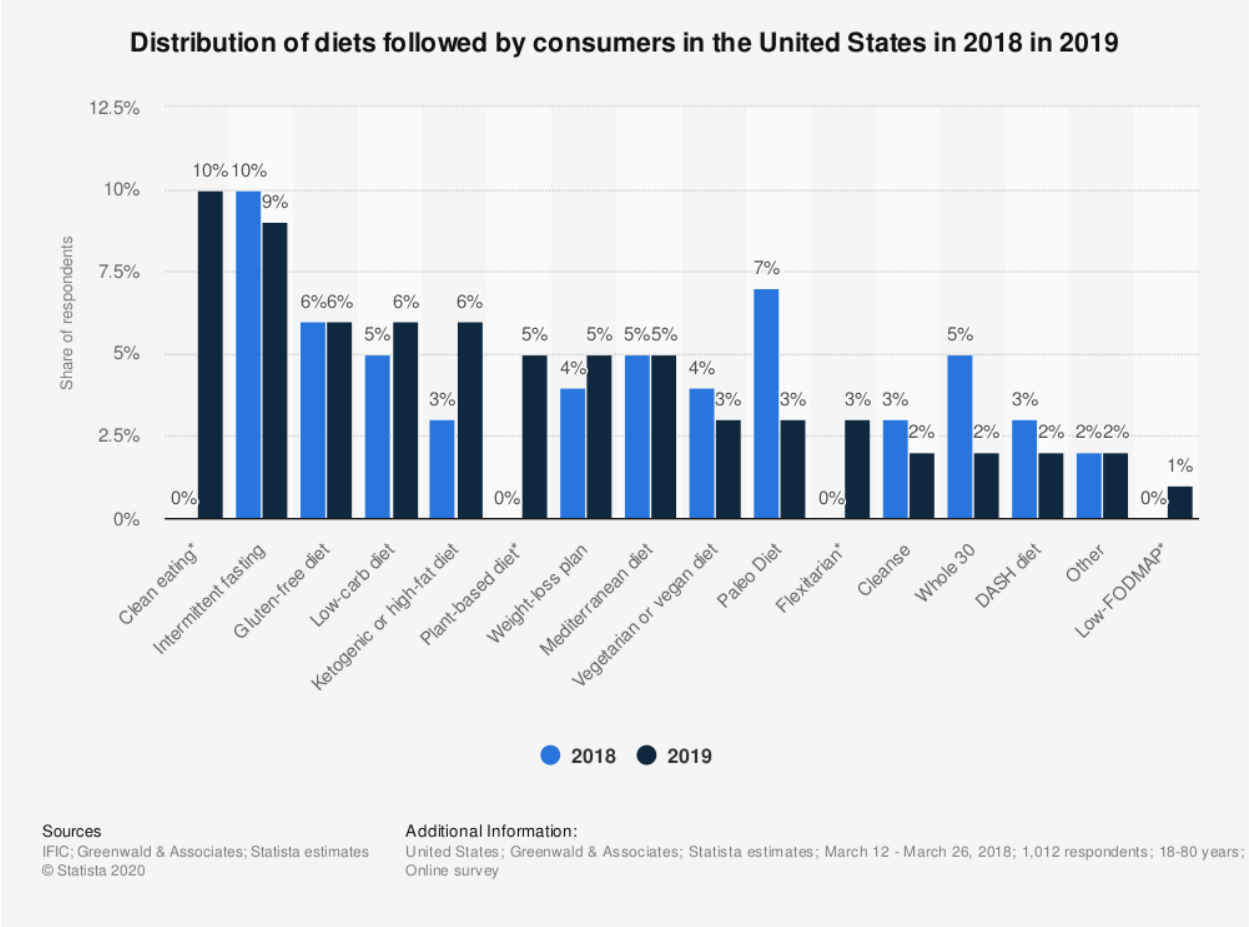
We all know the statistics but what these statistics translate in is that almost everyone now knows someone with diabetes. The chance of ourselves or someone in our life receiving a cancer diagnosis is higher than ever. In other words, behind the numbers, there is a devastating loss of quality of life and life itself.

Not surprisingly, this drives people to search for ways to improve their overall health and decrease their risk factors. No one wants to become a statistic and people turn to diets and supplements to improve their lives and health as well as for prevention. According to the NCHS Data Brief No. 399, from February 2021[7], in 2017-2018 over 57% of adults over 20 in the United States have used dietary supplements in the past 30 days. In contrast, the percentage of people using supplementation in 2007-2008 was 48.4%, and the dietary supplement use increased for respondents from all age groups. This trend was exacerbated by the onset of the COVID-19 pandemic, as shown in an online cross-sectional survey[8], with dramatic increase in supplementation in Asia (29.5% to 71.9% of respondents), America (40.6% to 75.7% of respondents), Europe (30.8% to 68.7% of respondents), and Turkey (21.3% to 62.2% of respondents).

Different diets are often marketed as the cure-all solutions in fighting excessive weight, insulin resistance, high cholesterol, and having a strong immune system. We have all heard of the benefits of a Ketogenic, Paleo, Vegan, Mediterranean and other diets, but oftentimes adhering to a strict protocol is also associated with health risks. According to Harvard Health Publishing[9],

[7] https://www.cdc.gov/nchs/products/databriefs/db399.htm
[8] https://academic.oup.com/cdn/article/5/Supplement_2/207/6293090
[9] https://www.health.harvard.edu/staying-healthy/should-you-try-the-keto-diet

keto diet comes with serious risks such as nutrient deficiency, kidney and liver problems, constipation, fuzzy thinking and mood swings.



This is just one example of the discrepancy of available information. For every study one finds arguing pro specific intervention, there is another one warning of the dangers. Some studies warn us that eating food high in saturated fat, especially coming from animal sources, could lead to ischaemic heart disease, high cholesterol and hypertension, while others claim no such relationship exists. Hearing contradicting information with no way to explore those relationships through data makes it harder to decide what nutrition fits our needs and preferences the most. The choice is even harder when food sensitivities and allergies are limiting our options. This

only shows that popular diets are not a one size fits all, and in order to be able to create an optimal nutrition plan for ourselves, we have to understand the core reasons behind a specific eating protocol, something we can do by exploring the relationships between food groups and health conditions.

With all those concerns in mind, while also questioning the origin of our food and being concerned about contamination or genetic modification we are starved for accurate, concise information. The increased accessibility and popularity of the Internet makes the task of finding and exchanging information much easier. Social media threads and groups are gaining popularity, and allowing people to connect, support each other, and share the knowledge they have obtained. However, online sources are often contradicting each other, and the same question has many different, mutually exclusive answers. Finally, the health and wellness industry is constantly promoting supplementation as the only way to get all of the vitamins and minerals our bodies need, a trend that started in the middle of the 20th century, and bloomed over the past decades. All of these factors, accompanied by the easy access to supplementation products, naturally leads to increasing use of dietary supplements in an attempt to increase our overall health as seen above.

Between the diets and the supplements, we are trying to take control over our health, attempting to reverse damage that has already been done, counter the effects of risk factors, avoid diseases and injuries, provide the optimal resources to our bodies to fight bacteria, viruses, and other offenders, and ensure we can have a longer, happier, and more productive life.

Is Food Making You Sick? aims to visualize the patterns and correlations between health and nutrition in an accessible way, in order to allow people to examine and analyze these relationships. The basis for the visualization is that before we can navigate the vast land of

information on the internet, we need to understand how food and nutrition relates to disease. We will present a series of visualizations that allow users to observe how a food/nutrient relates to a disease.

The main goal is to empower users to explore relationships they may not be aware of by juxtaposing data on nutrition consumption and health outcomes. In the process, we will answer questions such as do countries that have a high fat diet, also struggle with heart disease. Do countries consuming high amounts of sugar are overburdened with diabetes and vice versa?

Users will be able to analyze the correlations in isolation, explore potential patterns, and better inform their choices. Countries can be compared and trends and patterns related to geographic location and population size may be explored.

Is Food Making You Sick? utilizes available research and the data collected in the Global Dietary Database and the Global Burden of Diseases database to ensure the visualizations reflect the most extensive, up-to-date, and quality-checked information available. Through these sources, users can have access to a visual representation of large amounts of data and have the opportunity to explore a multitude of relationships in an intuitive, easy to grasp way.

Is Food Making You Sick? also aims to spark discussion in the science community and provoke research of less well-known relationships, which may be statistically significant, and examine more ways in which different food groups and types can improve or harm our health. Patterns and trends in different countries could provide valuable information about the ways in which traditions, access to specific foods, and local cuisine are related to a nation’s health and disease prevalence.

Methodology

Data

Data Sources

The project utilizes two different data sources: The Global Burden of Disease and Global Dietary Intake Estimates, both containing data for the years 1990, 1995, 2000, 2005, 2010, 2015 and 2018. These datasets, combined together provide the ability to create visualizations that allow users to observe and analyze correlations between 47 nutritional and over 90 health metrics in 184 countries.

Other sources were considered, such as the Global Nutrition Report, Food and Agriculture Organization of the United Nations (FAO) Supply Food database, WHO's Global Health Estimates, World Bank's Health Nutrition and Population Statistics, but the Global Burden of Diseases and Global Dietary Database were selected as they provide exhaustive data, spanning a long period of time (1990-2018), and went through a rigorous process of quality check by a consortium of scientists and researchers.

The Global Burden of Disease (GBD)

Institute for Health Metrics and Evaluation (IHME) at the University of Washington[10]

[10] https://ghdx.healthdata.org/gbd-results-tool

The Global Burden of Disease database provides information about the impact of more than 300 diseases, injuries, and risk factors in more than 200 countries and territories. This helps quantify the actual loss diseases and injuries cause - lives lost, years lost, years lived with disability, etc. The database contains more than one billion data points, available to the public. The data contain information by location, year, context, age, metric, measure, sex, and cause, and can be filtered out in the GBD website before download, as well as explored with the GBD visualization tool.

The database allows selection of different measures, from which deaths, incidence, prevalence, DALYs (disability-adjusted life year), YLDs (years lived with disability), YLL (years of life lost) were selected for initial exploration and data analysis. The exact measures to be included in the final project will be determined after exploratory analytics are run. These measures represent the many ways in which diseases and injuries affect people's lives as well as the economic consequences of diseases and disabilities. These measures are defined by the IHME[11] as follows:

- Incidence - the number of new cases of a given disease during a given period in a specified population. It also is used for the rate at which new events occur in a defined population. It is differentiated from prevalence, which refers to all cases, new or old, in the population at a given time
- Prevalence - the total number of cases of a given disease in a specified population at a designated time. It is differentiated from Incidence, which refers to the number of new cases in the population at a given time
- Disability-adjusted life years (DALYs) - the sum of years lost due to premature death (YLLs) and years lived with disability (YLDs). DALYs are also defined as years of healthy life lost

[11] https://www.healthdata.org/terms-defined

- Years lived with disability (YLDs) - years of life lived with any short-term or long-term health loss
- Years of life lost (YLLs) - years of life lost due to premature mortality

The metrics are the ways in which the measures - incidence, prevalence, deaths, DALYs, YLDs, and YLLs are measured and compared. From all available measures, number, percent and rate (per 100k people, dividing the deaths caused by the specific reason by the country's population) were selected for initial analysis. Both percent and rate normalize the data between countries to allow comparison, and number gives the absolute number of people affected by the condition without consideration of the total population of the country.

There are over 350 causes (disease, injury, risk) in the database, arranged hierarchically in four levels, increasing specificity with each level. The top level contains just four main groups, while the lowest, fourth level, represents diseases at the most detailed level. If any existing condition is not represented explicitly in a separate category on any of the levels in the hierarchical structure, it is added to “other” in the respective category of the same level. Due to this, all categories within a specific level are mutually exclusive and exhaustive, and all diseases, injuries and risk factors are counted exactly once.

The data sources used to create and underline the GBD database include administrative records, censuses, clinical trials, demographic surveillance, disease registries, environmental monitoring, financial records, surveys, vital registration, and more.[12]

[12] https://www.healthdata.org/sites/default/files/files/Projects/GBD/March2020_GBD%20Protocol_v4.pdf

Global Dietary Database - Intake Estimates (GDD)

The Global Nutrition and Policy Consortium is an initiative based at the Tufts Friedman School of Nutrition Science and Policy[13]

The data in the Global Dietary Intake Estimates database estimates the mean intake for each dietary factor included in the dataset by country, year, and subgroup, using Bayesian hierarchical prediction model. The data are derived from over 1200 survey-years of data from public and private sources, which have been reviewed, standardized, and approved for the GDD 2018 classification model inclusion. They have been categorized and harmonized to maximize comparability between the various surveys and countries. FoodEx2— "a sophisticated food description and classification system developed by the European Food Safety Authority (EFSA)" — is applied to the data in order to reduce assessment errors caused by self-reported food items. The database structures the estimates in 3 geographic or economic levels: global, superregion, and country.[14]

In addition to the food intake surveys, the prediction model is given a broad range of covariate data from diverse set of data sources, such as FAO food balance sheets data, 1980-2018, Harvard Global Expanded Nutrient Supply (GENuS) data, 1980-2013, World Bank GDP, 1980-2015, Precipitation, 1982-2014, Unemployment rate, 1991-2015, Education years, 1980-2010, Gini coefficient, 1980-2015, Poverty rate, 1991-2015, etc.

The GDD dataset estimates mean levels of dietary intake for over 50 dietary factors for 185 countries between 1990 and 2018. The data are separated in age groups, sex (male and female), residence (urban and rural), education level, and also provide totals for each group,

[13] https://www.globaldietarydatabase.org/
[14] https://www.globaldietarydatabase.org/methods/summary-methods-and-data-collection

such as total for all education levels. When covariate data were missing in the raw data used for the database, they were imputed using linear interpolation, moving average, or by assigning region-level means to countries.

The Global Dietary Intake Estimates dataset contains 144 CSV files with estimates on global, regional, and country-level for over 50 dietary factors. These dietary factors[15] belong to four major groups - foods, beverages, macronutrients, and micronutrients. They are as follows:

- **Foods**: Fruits (measured in g per day), Non-starchy Vegetables (measured in g per day), Potatoes (measured in g per day), Other Starchy Vegetables (measured in g per day), Beans and Legumes (measured in g per day), Nuts and Seeds (measured in g per day), Refined Grains (measured in g per day), Whole Grains (measured in g per day), Unprocessed Red Meats (measured in g per day), Total Processed Meats (measured in g per day), Total Seafoods (measured in g per day), Eggs (measured in g per day), Cheese (measured in g per day),Yogurt (including fermented milk) (measured in g per day)
- **Beverages**: Sugar-Sweetened Beverages (measured in g per day), Fruit Juices (measured in g per day), Coffee (measured in cups per day) (1 cup = 8 oz), Tea (measured in cups per day) (1 cup = 8 oz), Whole Fat Milk (measured in g per day), Reduced Fat Milk (measured in g per day), Total Milk (measured in g per day)
- **Macronutrients**: Total Carbohydrates (measured in % total kcal per day), Total Protein (measured in g per day), Total Animal Protein (measured in g per day), Plant Protein (measured in g per day), Saturated Fat (measured in % total kcal per day), Monounsaturated Fat (measured in % total kcal per day), Total Omega-6 Fatty Acids (measured in % total kcal per day), Seafood Omega-3 Fatty Acids (measured in mg per day), Plant Omega-3 Fatty Acids (measured in mg per day), Dietary Cholesterol

[15] https://www.globaldietarydatabase.org/GDD/VariableDefinitions

(measured in mg per day), Dietary Fiber (measured in g per day), Added Sugars (measured in % total kcal per day)

- **Micronutrients**: Dietary Calcium (measured in mg per day), Dietary Sodium (measured in mg per day), Iodine (measured in ug per day), Iron (measured in mg per day), Magnesium (measured in mg per day), Potassium (measured in mg per day), Selenium (measured in ug per day), Vitamin A with Supplements (measured in ug RAE*/day), Vitamin B1 (measured in mg per day), Vitamin B2 (measured in mg per day), Vitamin B3 (measured in mg per day), Vitamin B6 (measured in mg per day), Vitamin B9 (measured in ug per day), Vitamin B12 (measured in ug per day), Vitamin C (measured in mg per day), Vitamin D (measured in ug per day), Vitamin E (measured in mg per day), Zinc (measured in mg per day)

**RAE - retinol activity equivalents, or the amount of vitamin A that can be actively absorbed by the body.*

Data Manipulation

Is Food Making You Sick? utilizes the Global Burden of Diseases and the Global Dietary databases, which provide information on health and nutrition metrics respectively. Both datasets contain data from 1990 to 2018 on global and country level, and contain demographic information such as sex and age. Due to the extensive information the databases contain, and the fact that they are meant for consumption by scientists for further research, the size of the data files is bigger than what can be consumed by end users. Because of this, there were several steps involved in the preparation of the final dataset the project uses in order to distill the data to the most actionable and easy to interpret by users variables.

The Global Dietary Database dataset contains three sets of CSV files, each set currently having 47 files. Each of them represents one nutrition variable, for example Total processed meats, which is encoded numerically in the file name and the provided nomenclature document. In order to prepare the data for injection and analysis, a pipeline was created, which processes and merges each separate CSV file into a master data frame, containing the information for all nutrition variables on the desired geographic or economic level: country, supperregion, or global.

Each file has information about one variable only, for all years and contains demographic information about sex, age group, education level, and residency. Depending on the level, geographic or economic data is either omitted (for the global datasets), or specified with specific encoding. The seven superregions are East & Southeast Asia, Former Soviet Union, High-Income Countries, Latin America & Caribbean, Middle East & North Africa, South Asia, and Sub-Saharan Africa. At the most detailed level there are 185 countries, encoded in the datasets with their ISO* codes.

** ISO 3166-1 is a standard defining codes for the names of countries, dependent territories, and special areas of geographical interest. It is the first part of the ISO 3166 standard published by the International Organization for Standardization.*[16]

The first step in preparing the data for injection and analysis, was creating a dictionary containing the country name and ISO code to serve as a master data frame with which all files with nutrition variables were merged once they were processed. The ISO code was used to compare all separate datasets with the dictionary before taking any further steps in preparing the data, and investigate if all countries have diet estimations. This showed that three of the 188 countries - Andorra (AND), North Korea (PRK), Somalia (SOM) - do not have any data yet. Those countries were removed. In further iterations, they may be included back to the dictionary,

[16] https://en.wikipedia.org/wiki/ISO_3166-1

as the Global Dietary Database is supposed to be updated on a rolling base and may include data for the removed countries in the future.

Each of the 48 datasets contains the mean daily intake of the respective nutrition variable in GDD variable units[17], as well as upper and lower confidence intervals. The variable code is only indicated under the ‘varnum’ column and the name of the file, and the mean intake and uncertainty intervals columns are identically named for each variable. Before the datasets were merged, this had to be resolved in a way that would prevent overriding and losing data for any of the variables. To handle this, the numeric code of each nutrient, food or beverage observed, was appended to the beginning of the column names indicating mean intake and uncertainty intervals. Next, the original columns containing variable information, such as numeric code, type, and description, were removed from the dataset in order to reduce file size.

At this stage of the project, the goal is to visualize the estimates for all age groups, both sexes, all education levels, and residents of both urban and rural areas. The Global Dietary Database provides those estimates on a country level, encoding the respective rows with totals as 999. Taking advantage of this, all other rows, where the value of age, sex, education, or residency were not equal to 999, were removed to further reduce the files’ size and ensure better performance. Once the total estimates were filtered, the numeric value indicating them was no longer needed as information, and age, sex, education level, and residency columns were removed from each of the 47 datasets.

Due to differences in some of the nutrition variables and the data available for them, some datasets provide information on serving size and the respective serving size upper and lower

---

17 https://www.globaldietarydatabase.org/sites/default/files/available-for-download/2019-12/GDD%20variable%20definitions.pdf

uncertainty levels. To unify the shape of each dataset, those columns were dropped at this stage of the data manipulation process. All processed datasets were saved in new files, and the original datasets were preserved to ensure no data is lost, and can be used if and when the project needs to provide more detail from columns or rows that were removed at this stage. Finally, all datasets were merged with the original mapping of country names and ISO codes, resulting in a master data frame containing 1295 rows and 145 columns.

Next, the Global Burden of Diseases dataset was prepared for use. The Global Burden of Diseases database allows selection of eight different components of the data: location, year, context, age, metric, measure, sex, and cause. This provides a lot of flexibility in filtering the data, and the initial selection for the project was as follows:

- Location: included global and country level data, total of 205 locations
- Year: included the years available in the Global Dietary Database to allow comparison: 1990, 1995, 2000, 2005, 2010, 2015, and 2018
- Context: included only cause, as this is the most objective context from the provided selection, as well as easy to grasp by the general public
- Age: at this stage of the project, the selection is an estimate for all age groups, and age is not included in the analysis of correlations in order to provide a general overview and decrease the complexity for users. Instead, the provided total for all age groups is used
- Metric: number, percent, and rate were selected, as all three of them represent the data in a different way, in an absolute and normalized scale respectively
- Measure: Deaths, Incidence, Prevalence, Years of Life Lost, Disability-Adjusted Life Years, and Years Lived with Disability were selected to provide a more detailed overview of the actual effects diseases have on people’s lives
- Sex: a total estimate for both sexes was selected at this stage of the project

- Cause: at the initial stage of the project 98 diseases were selected in order to examine potential correlation and significance. After data analysis and exploration of the relationship between cause and nutrition, the number of causes could be reduced as needed in order to decrease data noise

The selected criteria yielded six CSV files, each containing half a million rows. As they were split based on the row count only, and not on specific criteria, such as year, metric, or measure, all files had to be appended. Before appending them, all six datasets were treated to reduce their size. First, unnecessary columns were removed: location_id as the final dataset will utilize country names and ISO codes, sex_id and sex_name as the selected data contains total estimates for both sexes, age_id and age_name as the data is a total estimate, and not separated by age groups.

Since the countries in this dataset are over two hundred, while the Global Dietary Database contains only 185, a cross-check was performed between the two datasets. Discrepancies in the spelling of locations were examined and all countries, a total of 20, were encoded as spelled in the Global Dietary Database. Countries which did not exist in both of the datasets simultaneously, were removed and the final number of locations having data for both nutrition and health metrics, for all selected years, is 184. Finally, the six separate CSV files containing all selected data from the Global Burden of Diseases database were appended into a master data frame containing over 2.3 million rows.

In order to enhance the analytical capabilities of the project, another layer of country metadata was added to the two main datasets. Data on population as of 2020 was obtained from the World Bank API, cleaned, and the 184 country names were synchronized with the original

country list. Underweight, overweight and obesity rates for 2016 (the latest available data) from the World Health Organization API were added to the population dataframe.

Data Analysis

The raw data from the Global Dietary Database and the Global Burden of Diseases contains approximately 6 Gigabytes of CSV files. They were cleaned, curated and prepared for analysis, resulting in over 3 million rows of data.

After the datasets were cleaned, they were analyzed using Python, Pandas and Matplotlib. A subset of the data was initially selected as a proof of concept, showing the incidence rate of Diabetes mellitus type 2 for 2018. Using this, a test set of nutrients was plotted in relation to the selected health condition, juxtaposing the nutrient consumption and disease incidence rate for all 184 countries. Additionally, a regression line showing the relationship between nutrient and condition was plotted on top of the correlation in order to explore the strength and direction of the relationship better. After analyzing the test subset, all nutrients were plotted against each of the conditions, resulting in 47 PDF files with small multiples of each correlation.

Further data exploration was performed by plotting the data in an interactive d3.js visualization. Utilizing d3.js's capabilities to expose data on hover or click, the next iteration of data exploration was performed on the frontend. After careful analysis of the available data, the 98 initially selected health conditions were then filtered out based on their potential to be influenced by nutrition. As a result, 14 of the original 98 diseases or conditions were removed from the final

data served to users. Examples of such conditions are transport injuries, interpersonal violence, or self-harm.

Implementation

Hosting

Is Food Making You Sick? is built by analyzing and extracting the datasets powering it, from almost 6 Gigabytes of raw data from two extensive static sources - the Global Dietary Database and the Global Burden of Diseases. As there is no existing API with these datasets, a project specific API was built, storing the data in a MySQL database in a shared hosting platform. The API layer is implemented in Node.js and Express and is hosted on Heroku, running the Node.js script continuously. The front end itself is built in d3.js, Vue, Vuex, and Vuetify, and the live application is hosted on GitHub pages.

Back end

The MySQL database is accessed through an API layer, with an endpoint taking the selected nutrient and condition as query parameters. It is built in Express and Node.js, and uses the memory-cache module to cache the data for one week,increasing performance and reducing costs. Based on the query parameters of the API call made on the front end with the selected nutrient and health condition, the Node.js/Express script queries the database. As part of the query, the three tables - nutrients, diseases, and metadata - are joined and the response is prepared for plotting on the frontend.

Database

After cleaning, formatting, and analyzing the nutrition and health data, the two master datasets were injected in two MySQL tables - one storing the nutrition data, and one storing the health conditions data. A third, supplementary table was created, storing country metadata such as population data for 2020 sourced from the World Bank API, obesity, overweight and underweight rates for 2016 (latest available data) sourced from the World Health Organization. These metadata are then used to provide an additional layer of information, complimentary to the two main data sources. On the frontend this is indicated by changing the size of the data points representing each country, scaling based on the selected metric.

Ingesting the database was performed with Sequel Pro, which allows direct upload of data stored in local CSV files and speeds up production time. After the data were prepared and the final CSV files holding nutrition, health, and meta data were exported with Python, they were imported into the respective tables in the database with the help of Sequel Pro.

The MySQL database is hosted on a shared hosting platform. A read-only user has access to the data through an API endpoint, querying and joining the three tables based on the query parameters selected by the user - id of the nutrient and id of the disease. As part of the query, some variables such as the value for the selected nutrient and the prevalence rate for the selected disease, were type-casted as floats, which could not be executed during ingestion due to some limitations of the available data types in Sequel Pro. This way the data are served to the front end in the proper format, ready to be plotted with d3.js.

Front end

The front end of the application is built in d3.js, Vue, Vuex, and Vuetify, and hosted with GitHub pages. Based on users' selection of a nutrition and a health condition, the application is making an API request to the Heroku Node.js API endpoint, which queries the MySQL database, hosted in a shared hosting server, joins the three tables, and returns the results to the front end in a structured, ready to visualize way.

User Experience and User Interface

The user interface contains three main views - two introductory pages which show how the data are distributed for nutrition and health condition respectively, and the main visualization, showing the relationship between the selected nutrient and selected disease. Due to the complexity of the project, creating a seamless user experience is crucial for engaging the audience and ensuring users can benefit from exploring and analyzing the visualized relationships between nutrition and health. This problem is addressed by the two introductory pages, following a case study which explores the relationship between total milk consumption and prevalence of rheumatoid arthritis per hundred thousand people in each of the 184 countries.

Is Food Making You Sick? takes users through a short journey, introducing the audience to the building blocks of the data in the first two views of the application. It takes total milk consumption (in grams per day) and rheumatoid arthritis prevalence per 100 thousand as a case study to show what stands behind the main visualization, how the data are distributed, and how the correlation is made. This part of the project aims to introduce users to the concept of

the application and familiarize them with the data in a step by step manner. The user interface on the first two pages does not allow any changes of the selections or filtering, as it presents viewers with preselected values for the nutrient and disease to create a case study, and is simplified in order to allow users to give all of their attention to the underlying data. Users can introduce changes and select any of the almost four thousand relationships on the third, main view of the application. This is also the view that allows correlations to be seen and explored, and countries to be compared based on the continent they are located in, the nutrient consumption, disease prevalence, and an additional, complimentary metric such as population, obesity rate, overweight rate, and underweight rate.

When landing on the project page, users are introduced to the total milk consumption in the 184 countries explored in the application. The graph has only a vertical Y axis, showing the average daily consumption in grams, which naturally orders the countries from highest to lowest milk consumption. A tooltip exposes the exact values and names of the country users hover over, allowing access to the concrete data behind each point. Additionally, there is a separate tooltip containing the same information for the United States of America, which is permanently exposed, so users can quickly orient themselves.

The circle representing the United States of America is highlighted in red at all times, showing the ranking of the country in a visual, easy to digest way. This way users can intuitively understand the approximate location of the country, for example above, below, or close to the median, as the scale is continuous. Observing the data in this way also shows the distribution of the data - seeing that most countries are actually consuming less than 250 grams of milk per day. Additionally, outliers, such as Lithuania, which has average daily milk consumption of 631.2 grams, are easily visible, and the distance between data points shows the disparity of the data. Finally, as the data points are plotted on top of each other, which occurs when countries have

the same or similar score, the circles look darker, emphasizing the density at some points of the visualization.

Proceeding to the next view, users are exposed to the available information about prevalence rate (amount of people currently living with a certain condition per hundred thousand) of rheumatoid arthritis in all 184 countries. This page is using the same minimalistic visual language, utilizing one vertical Y axis which sorts the countries from highest to lowest rate, and a tooltip, showing the information belonging to each data point on hover. Again, the information about the United States of America is exposed permanently, indicating the relative ranking of the country compared to all others. In this view, the tooltip shows not only the country name and the prevalence rate of the disease, but also approximately how many people in the specific country are affected by the condition. Both values are present for their respective reasons - the rate is normalized and not influenced by the population and can be used to rank and compare the countries, and the approximate number of people is giving a perspective on the actual impact of the disease on people.

The final view, which is also the core component of the project, finalizes the case study presented to users. It shows the correlation of total milk, as an example nutrient, and rheumatoid arthritis, as an example disease, in an interactive scatter plot. Exploration and analysis are enhanced by a multitude of interactions, enabling different aspects of the available data to be seen or highlighted.

Each interaction slightly changes the view, altering the way data is displayed based on users' selection. It incorporates two hovers on different elements, four drop down menu selections, a selection of data points visibility based on selected region, color coding, and the plotting of a regression line, showing the strength and direction of the relationship. Due to the distribution

and density of the data in some areas of the scatter plot, a zoom functionality enables users to zoom into a specific part of the visualization and explore the countries in the visible range. (All interactions are reviewed in detail in the *Interactions* section). The selection of the data and interactions showing different aspects of it are possible through just several user interface components, enabling users to make their choices in a simple step by step process.

At the top of the page are the two main dropdowns - one to select a nutrient and one to select a health condition respectively. They are the main tool users have to change the displayed data and explore the relationships between the 47 nutrients and 84 diseases available at their disposal. Making a change to any of the two main dropdowns would trigger an Ajax call to the API running on Heroku, and query the MySQL database, providing the data for the requested relationship. The returned data is then displayed on the main visualization, changing the position of each country on the bottom X or left Y axis depending on whether users selected a new nutrient or a new disease. To ensure the user experience is clear and intuitive, a progress circle is displayed on top of the visualization while the data are loading, indicating the latter to users. Once the server returns the data for the requested relationship, the circles representing countries animate, changing their positions on the scatterplot.

Below the selection menus is the title of the visualization, which is getting dynamically updated based on the selected nutrient and disease in the dropdowns above it. It provides a simple sentence “Relationship between *selected-nutrient-name* and *selected-disease-name*”.

The last major component of the user interface is the side selection panel, which encompasses the rest of the user interface elements related to the way users interact with the visualization and explore the data. It holds controls which alter the metadata displayed and the way data points are styled in order to indicate different aspects of the underlying data. Users can make

selections through two dropdown menus highlighting a country on the scatter plot, and changing the size of the circles based on a selected scale respectively, an interactive legend allowing selection or deselection of a region or continent, and a slider plotting or hiding the regression line indicating the strength and direction of the relationship between nutrient and disease.

Users have the ability to scale the circles representing the 184 countries based on four different scales - population, obesity rate, overweight rate, and underweight rate. The radius of each circle is calculated based on the value the country holds for the selected measure - the bigger the circle is, the higher the score of the country. For example, when population is selected for sizing, China and India will have the biggest circles and will be much more visible than other countries, while circles representing countries like Iceland or Saint Vincent and the Grenadines, which have populations of less than half a million people, will be almost invisible. When underweight rate is selected, most countries' circles are scaled down to almost being invisible, and countries located in Asia and Africa are visually dominating the visualization, due to the higher rates of undernourishment in parts of these two continents.

The interactive legend is showing the color coding based on the continent each country belongs to. It allows users to remove the color coding of a selected region, or look at one region in isolation. *Interactions explained in detail in the Interactions section.*

Visualization

The main visualization - the correlation scatterplot - plots the nutrient consumption data on the bottom X axis, and the disease prevalence rate (amount of people living with a specific health condition per 100 thousand people) on the left Y axis.

The X axis is labeled with the nutrient name, the food group it belongs to, and the measurement used to compare the countries. For example, the first time users get to the main visualization, the preselected milk consumption would display the following: “Total Milk - Beverages (grams per day)”. All data points are plotted horizontally based on how they scale on the nutrient consumption, and countries scoring highest on this measure will be at the far right corner of the visualization.

The selected health condition is represented on the left Y axis. It shows the name of the condition and the measure - prevalence per 100 thousand. For example, when initially exploring this visualization, the axis shows “Rheumatoid arthritis - Prevalence per 100k”. All data points are placed vertically based on the rate of the selected disease, and countries scoring the highest will be located at the highest point of the visualization.

184 circles are plotted on the scatter plot at all times, each of them representing one country, located at the exact point representing the score on the X axis and Y axis simultaneously. The circles initially have the same radius, but sizing can be augmented based on user selection. All countries are color coded based on their geographic location, representing the continent they are located in. This can also be augmented by users by selecting or deselecting a continent from the side menu.

A circle representing one specific country can be highlighted in two different ways: first, by using the ‘highlight country’ drop down menu - when a country is selected, the respective circle border gets thicker and changes its color to black, keeping the inner part of the circle the color representing the continent the country is located in. Secondly, when users hover into a specific circle in the visualization, its border also gets thicker, and highlights in bright red. This way there

is a distinct visual difference between the two types of selection, and users are given an indication of what they are currently exploring or comparing.

The visualization has another component, which can be turned on and off based on user selection - a regression line calculated and plotted on top of the selected relationship between nutrient and health condition. This line indicates the strength and direction of the relationship. The steeper the line is, the stronger the relationship between the two variables. The direction of the line - going upwards from the bottom left to the top right corner, or going downwards from the top left to the bottom right corner - indicates the direction of the relationship. If the line is going upwards, this means that the relationship is positive, and if it is going downwards, the relationship is negative.

As we are investigating how food could affect health, the independent variable is the selected nutrient, and the dependent variable is the selected disease. In cases where the relationship is positive, such as the correlation between total milk consumption and rheumatoid arthritis prevalence per 100 thousand people, this means that increase in consumption of the selected nutrient is related to increase in the prevalence of the selected disease. In cases where the regression line is going downwards, indicating a negative relationship, such as the correlation between total milk consumption and prevalence of nutritional deficiencies per 100 thousand people, this means that increasing consumption of the selected nutrient is related to decrease in the prevalence of the selected health condition.

Plotting the regression line and visually indicating the strength and direction of the relationship is a crucial part of the visualization, as advanced statistical knowledge and experience in analyzing visual correlations is sometimes needed to observe the trend and pattern in the data.

Removing this burden from the users allows better reading, exploration, and analysis of the data.

Interactions

Is Food Making You Sick? aims to provide users with the ability to explore and analyze a great amount of data seamlessly. In order to achieve this, the project heavily utilizes interactivity, ensuring users can unearth different aspects of the data in an intuitive way, removing any visual noise or cognitive burden caused by the sheer amount of data incorporated in the project.

The legend on the right side of the visualization includes not just the visual explanation of what is the meaning of the different colors of the circles, but also allows users to filter and highlight the data. By default all countries are color coded based on the continent they are located in. Selecting or deselecting a checkbox from the legend would result in changing the color of the circles representing countries in the deselected continent to gray, and lowering their opacity. This is making those data points less visible and pushing them to the background, while also keeping them on the visualization, so users can continue exploring the selection they made in the context of all 184 countries. This allows analysis and exploration on a detailed level, while keeping the important context of the global distribution. If color coding countries by continent is not desired by the user at a specific point, they can deselect all continents, and all countries will be equally visible, taking a neutral gray color.

Users can also hover over the legend to explore the countries in a specific continent. This interaction is only available when the continent is selected, so all deselected regions cannot be

highlighted until they are selected again. When taking advantage of this functionality, users can see only the continent of interest color coded as per the legend, and all countries in other continents are again pushed to the background with decreased opacity and neutral gray color. This allows exploration of the distribution of a specific region while again keeping the context of the global distribution. For example, users can see how countries in Europe are distributed - both their proximity to each other and their relative position in the global context.

Users can also highlight a specific country from the ‘highlight country’ drop down menu. This highlights the circle representing the selected country by giving it a thick, black border and higher, constant opacity. If the country is located in a region that has been deselected in the legend, it will be colored in a darker shade of gray, keeping a visual cue so users can compare it with other regions and the global distribution of the data points. This allows comparison between a specific country and other countries, regions, or the global distribution, even if the region it is located in is not of interest and has been deselected.

Another interaction users can utilize is highlighting a circle representing a country by hovering over it directly in the visualization. This interaction is separate from highlighting a country from the drop down menu, and the visual indication for it is also distinct. When a circle is hovered, it gets a bright red thick border, even if the region in which the country is located has been deselected. Additionally, when a circle is hovered over, a tooltip is displayed, indicating the name of the country, its population, consumption of the selected nutrient and the measurement used for it, prevalence rate of the selected disease, and the approximate number of people affected by the condition. This number is calculated based on the population and the rate of the disease.

As some datasets are densely distributed at certain points, a zoom functionality is implemented. This way users can navigate to a specific point of interest and explore all countries easily, hovering over each of them. This resolves any issues caused by overplotting dense data and in combination with the sensitive tooltip, ensures visibility and representation of each country.

Findings

Is Food Making You Sick? is exploring the relationship between 47 nutrients and 84 health conditions, resulting in close to four thousand distinct relationships. Although correlation does not mean causation, it is important to study the relationships between food and disease.

The project's main visualization is utilizing a regression line in order to aid the exploration of the correlation. It allows users to quickly see if a relationship exists between the two selected variables, and how strong this relationship is. If the line is horizontal, this means there is no relationship, and the steeper the slope is, the stronger the relationship between the selected nutrient and health condition.

Additionally, the direction of the regression line shows the direction of this relationship. If the regression line is going upwards, or from the bottom left to the upper right corner of the visualization, the relationship is positive. Positive relationships exist when increase in one of the variables is related to increase in the other variable as well. If the line is going downwards, or from the top left corner to the bottom right corner of the visualization, the relationship is negative - increase in one of the variables is related to decrease in the other one.

In the presented case study - relationship between total milk consumption and prevalence rate of rheumatoid arthritis - we establish that in the data used for the application, there is an underlying strong, positive relationship between the two variables, meaning that higher milk consumption is related to higher rheumatoid arthritis prevalence. This is clearly visible from the slope and direction of the regression line - it takes close to a 45 degree angle from the bottom left to the top right of the visualization.

Exploring other correlations between milk consumption and health conditions allows us to clearly see that some diseases are related to milk in vastly different ways. For example, when choosing cirrhosis and other chronic liver diseases due to NAFLD (nonalcoholic fatty liver disease), we see that the regression line is not as steep, and is closer to being horizontal. This means that the relationship between the two is much weaker.

Taking another health condition as example - nutritional deficiencies - we actually see a negative relationship. This means that higher milk consumption is related to lower prevalence of nutritional deficiencies.

Milk is sometimes also linked with increased risk of cancer, however, the evidence is mixed, and inconclusive. Looking at the raw data - exploring the relationship between milk and several types of cancer, we can see that the relationship is vastly different, depending on the type of cancer we are looking at in relation to milk consumption. For example, Non-Hodgkin lymphoma, breast cancer, ovarian cancer, and kidney cancer have a strong, positive relationship to milk consumption. This means that as total milk consumption, so does the prevalence of those types of cancers. On the other hand, liver cancer or gallbladder and biliary tract cancer have a very weak positive relationship to milk, and even though the data show that the prevalence of those diseases increases with the increase of milk consumption, the change is not as drastic as it is

with the above group, where the relationship is strong. And finally, diseases such as cervical cancer have a completely different relationship to milk, which is slightly negative. This is again observed by plotting the regression line, which is almost horizontal, but slightly tilted downwards. This indicates there is either a very weak negative relationship - more milk consumption is associated with fewer cases of cervical cancer, and the prevalence of the disease decreases very slightly for each additional consumed gram of milk - or there is no relationship between the two.

An interesting finding is the correlation between added sugars consumption and prevalence of diabetes mellitus type 2. Although this relationship is oftentimes pointed as strong, with the help of the regression line, we can see that the actual relationship in the underlying data is not as strong. Exploring what other nutrients may affect the disease, we unearth that the correlation between dietary cholesterol consumption and diabetes type 2 is stronger than the one between added sugar consumption and diabetes type 2 - the regression line is much steeper.

Diabetes type 1, also known as juvenile diabetes or insulin-dependent diabetes, is a chronic genetic disorder, in which the body is attacking its own cells. In this way it is very different from diabetes type 2, which develops over time and is often linked to dietary factors. Because of this there is no expectation that sugar will be related to diabetes type 1 the same way it is related to diabetes type 2. When plotting the correlation and the regression line, we see that this is confirmed by the data - the regression line is almost perfectly horizontal, indicating there is no relationship between sugar consumption and diabetes type 1.

Exploring the relationship between dietary fiber and atopic dermatitis for example, the line shows a negative relationship - countries in which people implement more dietary fiber into their diet, also have fewer cases of atopic dermatitis.

The above mentioned examples of findings are only a few relationships. However, *Is Food Making You Sick?* allows the exploration of almost four thousand relationships. Using this powerful tool to explore the underlying data in a highly analytical way is an important data point in learning more about what are the ways in which the food we consume may influence our health. It can be used both by lay people in helping reduce their risk factors through nutrition and better food choices, but also by professionals in exploring and testing what we know about the correlation between diet and disease.

It is also an important first step in learning more about less researched nutrients, diseases, or relationships, or learning more about what food influences very well known, oftentimes very common diseases. Some relationships are very well studied and researched, and their outcomes are, to an extent, expected. The data do confirm this, as it is in the case of sugar consumption and diabetes type 1 - we do not expect sugar to have influence over diseases which are genetic, autoimmune, or caused by a viral infection. Other relationships however, are more surprising, making us question if we know enough, or even, if what we know is factual. Although the correlation between the two does not mean that the consumption of this particular nutrient is causing this disease, as human bodies work in much more complex ways and many factors may come into play when it comes to health, it is important to explore those relationships. On one hand, if the data we have collected is not reflecting the reality we are living, this should prompt us to look for better ways to collect, analyze and provide these data to the public. And on the other hand, exploring, exposing, and analyzing some relationships, questioning the results and our knowledge, can propel us to learn more. If the science is not conclusive, we should be eager to learn more, to probe the facts we have accepted as final.

Is Food Making You Sick? aims to explore the underlying relationships between nutrition and health, sparking a conversation about what makes us healthy and what makes us sick. Using the findings we get by exploring these data, we could strive to better our knowledge of the correlation between food and health, promote better dietary choices, and, whenever possible, use food as prevention of some health conditions. If this is achieved, we could have the power to reduce the devastating burden of some diseases - lives lost, years lost, quality time and productivity lost, and the very personal, oftentimes painful suffering they cause. By learning more and asking more questions, we can focus on actions that will lead to better life for many people.

Conclusion

Although correlation does not mean causation, it is extremely important to study and understand the relationships in the underlying data. This is especially crucial when it comes to the relationships between nutrition and health, as unearthing strong correlations may aid existing research or prompt researchers to explore new hypotheses. Even more important than that, it enables everyone, from the lay person to highly educated people, to have actual, high quality data in mind when making decisions about their own personal dietary choices.

Is Food Making You Sick? allows users to study close to four thousand total relationships between various nutrients - foods, beverages, macronutrients, and micronutrients - and 84 health conditions. It challenges the way we think about our health, by giving access to an extensive amount of both nutrition and health data, juxtaposing the two to reveal the strength and direction of the relationship between them. It also empowers researchers and health professionals to easily explore large amounts of data in an intuitive way, utilizing both the

simplicity of well known visualization types and the power of filtering, highlighting and interacting with the data intuitively.

Bibliography

1. World Health Organization. "The top 10 causes of death", December 9th, 2020. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.
2. Centers for Disease Control and Prevention. "Long-term Trends in Diabetes". April 2017. https://www.cdc.gov/diabetes/statistics/slides/long_term_trends.pdf.
3. Centers for Disease Control and Prevention. "Leading Causes of Death, 1900-1998". Accessed March 15, 2022. https://www.cdc.gov/nchs/data/dvs/lead1900_98.pdf.
4. The American Cancer Society. "Effects of Diet and Physical Activity on Risks for Certain Cancers". Last Revised: June 9, 2020. https://www.cancer.org/healthy/eat-healthy-get-active/acs-guidelines-nutrition-physical-activity-cancer-prevention/diet-and-activity.html.
5. Harvard Health Publishing, Harvard Medical School. "Should you try the keto diet?". August 31, 2020. https://www.health.harvard.edu/staying-healthy/should-you-try-the-keto-diet.
6. Suruchi Mishra, Ph.D., Bryan Stierman, M.D., M.P.H., Jaime J. Gahche, Ph.D., M.P.H., and Nancy Potischman, Ph.D. "Dietary Supplement Use Among Adults: United States, 2017–2018". Centers for Disease Control and Prevention, National Center for Health Statistics. NCHS Data Brief No. 399, February 2021. https://www.cdc.gov/nchs/products/databriefs/db399.htm.
7. Elif Aysin, Murat Urhan, Dramatic Increase in Dietary Supplement Use During Covid-19, Current Developments in Nutrition, Volume 5, Issue Supplement_2, June 2021, Page 207, https://doi.org/10.1093/cdn/nzab029_008.
8. Institute for Health Metrics and Evaluation. "Protocol for the Global Burden of Diseases, Injuries, and Risk Factors Study (GBD)". Version 4.0; Issued March 2020. https://www.healthdata.org/sites/default/files/files/Projects/GBD/March2020_GBD%20Protocol_v4.pdf.