

CSE 163 Intermediate Data Program

Final Project Proposal

Diet, Death, and Disease:

A Thorough Implication from Data Science

Author:

Todd (Chaoyuan) Zhang

Jasper (Jiexiao) Xu

Cynthia (Yutong) Pan

SECTIONS

	PAGE
Summary of Research	2
Motivation.....	3
Dataset.....	4
Challenge Goal.....	9
Method.....	10
Work Plan.....	12
Prof of Set Up.....	12

Summary of Research Questions:

1. **What is the inclination toward food in each country?** The variation in history, climate, and culture could lead to different dieting habits for each region. We are eager to know what is the eating composition of each country.
2. **Is there a correlation between the disease and diet?** The prevalent disease may have various causes such as genetics or particular professions. The application may let us know about the relationship between the food and common disease
3. **How could we improve our health through eating?** Some regions may prefer vegetables over meat, how could they improve to optimize the health condition in the long term perspective.

Motivation:

Diet: Diet is an indispensable part of a topic when it comes to the discussion of cultures and welfare. The analysis of the diet structure of different regions and groups allows us to further research related to history, public health, etc. It's possible to establish the connection between the diet and common diseases in the local area. While maintaining the diversity of diet in the light of cultural differences, people could moderately adjust their diet for optimal health conditions.

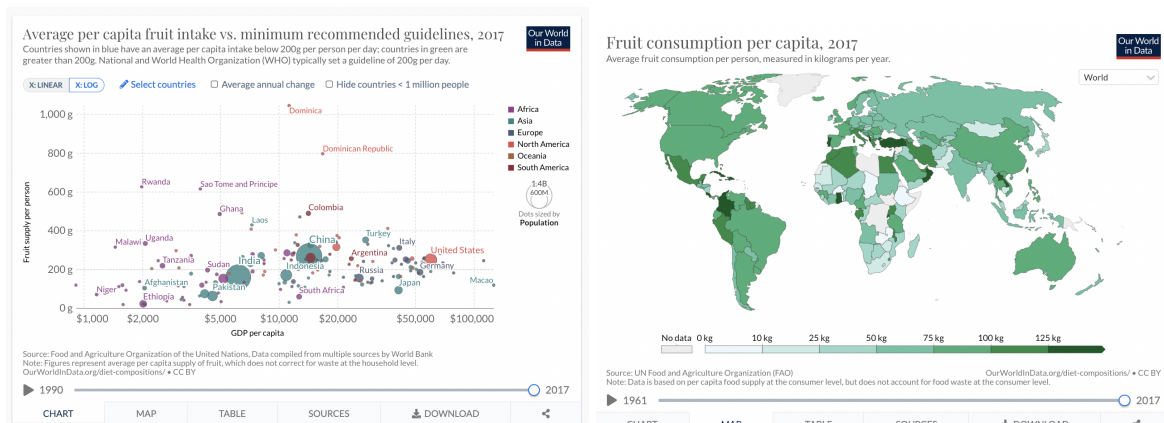
Dataset:

Diet Structure

1. Diet Composition

<https://ourworldindata.org/diet-compositions>

The Diet Composition Dataset comprises several sub-dataset in various aspects such as the “global vegetable consumption”, “vegetable consumption and guideline comparisons” and so on. Though each single dataset isn’t quite comprehensive and doesn’t embody the whole content of the current nutritional situation, they cover global data on a specific topic and have fine division of related factors under the given topic such as “vegetable supply per person” and “GDP per capita” in the “consumption vs recommend guideline” topic. Also, a merit of this data set is that it not only provides us with the statistical data in the format of csv, but also shows us a visualization of the data through the geographical data-map, helping us to drop attention on the important part quickly.



2. Disease Distribution

<https://ourworldindata.org/burden-of-disease>

The link attached here exhibits the data in the same way as the last dataset does. Differently, this one focuses more on the link between the diet structure and potential disease risk. Using both mortality and morbidity to evaluate the health outcomes in the current society, a few main corresponding factors leading to the diseases are involved in this dataset. Through categorizing and visualizing the data given, we can apply the divided data to machine learning and get a total estimation of the relationship between food, and go a step further to make judgments about the dietary problems that lead to disease. Through the prediction results, suggestions and adjustments can be made to perfect the diet structure, avoid potential problems and decrease the possibility of worsening disease burden.

Country	DALYs (Disability-Adjusted Life Years) - All causes - Sex: Both - Age: Age-standardized (Rate) DALYs per 100,000			
	1990	2019	Absolute Change	Relative Change
Afghanistan	86,375.17	55,424.65	-30,950.52	-36%
African Region (WHO)	81,293.92	50,162.52	-31,131.39	-38%
Albania	32,964.61	22,815.38	-10,149.23	-31%
Algeria	44,481.53	27,001.21	-17,480.32	-39%
American Samoa	35,230.49	32,712.13	-2,518.36	-7%
Andorra	22,259.04	18,988.56	-3,270.49	-15%

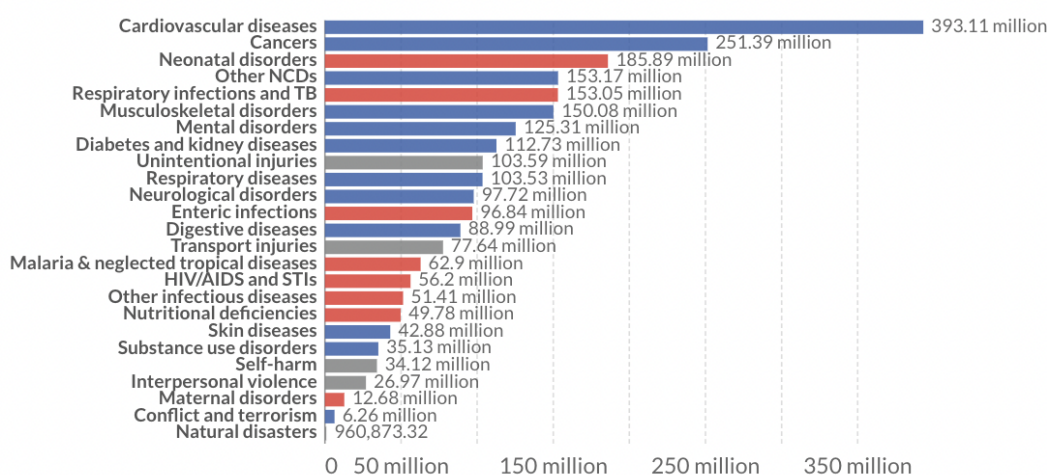
Burden of disease by cause, World, 2019

Total disease burden, measured in Disability-Adjusted Life Years (DALYs) by sub-category of disease or injury.

DALYs measure the total burden of disease – both from years of life lost due to premature death and years lived with a disability. One DALY equals one lost year of healthy life.

Our World
in Data

↔ Change country



3. Vitamin and Mineral Nutrition Information System (VMNIS) from WHO

<https://www.who.int/teams/nutrition-and-food-safety/databases/vitamin-and-mineral-nutrition-information-system>

This is a database from World Health Organization collecting the data connecting the “vitamin”, “mineral nutrition” and human health together. Information on different biochemical indicators are provided here for assessing the prevalence of various vitamin and mineral deficiencies in populations. Using this dataset, we can analyze the relationships between the vitamin, microelement and the distribution of the world human, based on which we can make a recommendation system on how/what to do with the current human nutrition situation, while also providing detailed advice on it.

<https://www.who.int/data/gho>

Mean BMI (kg/m²) (age-standardized estimate)

FILTERS

Last updated: 2017-09-27

EXPORT DATA in CSV format: Right-click here & Save link

Indicator	Mean BMI (kg/m ²) (age-standardized estimate)							
	2016			2015			2014	
	Both sexes	Male	Female	Both sexes	Male	Female	Both sexes	Male
	18+ years	18+ years	18+ years	18+ years	18+ years	18+ years	18+ years	18+ years
Afghanistan	23.4 [22 – 24.8]	22.6 [20.1 – 25.1]	24.1 [23 – 25.3]	23.3 [21.9 – 24.6]	22.5 [20.1 – 25]	24 [22.9 – 25.1]	23.2 [21.8 – 24.5]	22.4 [21.1 – 23.7]
Albania	26.7 [25.8 – 27.5]	27 [25.8 – 28.2]	26.3 [25 – 27.6]	26.6 [25.8 – 27.4]	26.9 [25.8 – 28]	26.2 [25 – 27.4]	26.5 [25.8 – 27.2]	26.8 [26.1 – 27.5]
Algeria	25.5 [24.5 – 26.5]	24.7 [23.4 – 26.1]	26.4 [24.9 – 27.8]	25.5 [24.5 – 26.4]	24.6 [23.4 – 25.8]	26.3 [25 – 27.7]	25.4 [24.5 – 26.2]	24.6 [23.5 – 25.7]
Andorra	26.7 [24.6 – 28.7]	27.3 [24.8 – 29.8]	26.1 [22.8 – 29.5]	26.7 [24.7 – 28.7]	27.3 [24.9 – 29.7]	26.1 [22.9 – 29.4]	26.7 [24.7 – 28.7]	27.3 [25.3 – 29.3]
Angola	23.3 [21.2 – 25.6]	22.3 [19.7 – 25]	24.3 [20.9 – 27.7]	23.2 [21.1 – 25.4]	22.3 [19.7 – 24.9]	24.1 [20.9 – 27.5]	23.2 [21.1 – 25.3]	22.2 [20.2 – 24.2]
Antigua and Barbuda	26.7 [24.6 – 28.8]	25.7 [23.2 – 28.2]	27.7 [24.4 – 31]	26.6 [24.6 – 28.7]	25.6 [23.2 – 28.1]	27.6 [24.3 – 30.8]	26.5 [24.5 – 28.5]	25.5 [23.5 – 27.5]
Argentina	27.7 [26.8 – 28.6]	27.8 [26.6 – 29]	27.6 [26.3 – 28.8]	27.6 [26.8 – 28.4]	27.7 [26.6 – 28.8]	27.5 [26.3 – 28.6]	27.5 [26.7 – 28.2]	27.6 [26.8 – 28.4]

4. Food Systems Dashboard

<https://foodsystmsdashboard.org/>

The *Food Systems Dashboard* not only provides the basic data that we would need for the analyze, but also the related policy and the recommended nutrition structure which we might need in deeper research. It's is not only a dataset, but some reasoning about the policy actions are introduced here, which might be a strong helper resource when analyzing the data we get.

One extra advantage that worths being mentioned here is that this dataset includes fine data and the select function, through which we can preview the data we need in several regions and then download to avoid the huge use of computer internal storage.

Agricultural actions		
	Action	What impact could the action have?
1	Deliver agricultural extension programmes, infrastructure and education to support farmers to grow and market nutritious foods	Increase availability and affordability of nutritious foods to local populations
2	(Re)design agricultural development programmes intended to increase food producers' income to also focus on producing, and accessing markets for, nutritious crops and providing nutrition education	Increase availability and affordability of nutritious foods to local populations

Challenge Goals:

1. Result Validity → How to ensure the regional difference in dieting habits is the main factor of common diseases?

One significant challenge for the analysis of disease is the variability of each subject. Diseases usually appear to correlate with the environment and genetics with which diet is a less conspicuous part. Further, the habit of eating can contribute to the formation of disease. Say, people who prefer having raw food are comparatively more likely to suffer from foodborne illness. If we take the aforementioned elements into consideration, the complexity of our analysis will rise significantly.

However, I believe the application of the test dataset will help us to prove the precision of our result. By dividing the dataset into two parts with a ratio of 7:3. We could examine the quality of the algorithm by the operation Accuracy Score and Squared Mean Error. (Machine Learning Goal) With such an approach, we could then know the role of food in people's health. Plus, statistical tools such as the p-value of the null hypothesis can also help us to figure out the robustness of the outcomes. In this context, we will set the alpha value to be 0.1, and our null hypothesis is that "the dieting habit is unrelated to the common disease."

Such challenges can happen in other research problems. The government policy could either foster the diversity of the minority such as LGBTQ people or smother them from thriving. Our predictions are inevitably susceptible to some factors but also have strong performance

2. Multiple Datasets/ Messy Data→ How do connect different datasets from different backgrounds and Years?

This is the greatest limitation of our project. As the time given for the project is limited, we don't have enough time to collect our own data through the survey and funding to obtain the data globally. Therefore, we can only utilize the data from the search engine. The data collected by the different organizations may have their own criteria for evaluation. The second-hand information produces the problem of generating our own index. Specifically, we have the index of LGBT from one website while another set provides data for gays, so it's not hard to predict that there will be an inconsistency between the dataset.

To overcome this difficulty, we could first think about using the mean() method to reduce the difference between each index. However, the index between two datasets is based on completely evaluating the system. The more feasible solution lies in that we need to have more datasets to resolve the major difference such as finding the dataset concerning the LGBTQ at the continent level, then applying the techniques from the class to clean the data.

Method:

1. **Collecting the data meets our research goals.** We could collect the geo or tabular data of **regional dieting**, and build the dataset between frequent diseases and diet structure.
 - a. There are three kinds of data very important in the dataset, the average caperal fruit, vegetable, animal product consumption in each geographical region, the related high-risk disease based on those areas, the food production, supply fact and policies in corresponding regions.
2. **Categorizing the data collected and sorting out the data needed.** After getting the dataset needed ready, we go ahead by grouping the data into the corresponding aspects that we need for further research.
 - a. For example, if we need to conclude the reasoning behind the diet structure and local high-risk disease, we might consider categorizing the diseases by their causes such as the deficiency in vitamin C, or the shortage in vegetables, etc.
3. **Processing the data based on the calculating system we choose.** Having the groups of data that were classified by the tags we need, we start to process the data so that they can have the same range and index setting background, avoiding the mess caused by using multiple datasets. This step also satisfies our second challenge goal.
 - a. For instance, if the dataset that we choose are different, then the following circumstance might happen: one dataset sets the range “0-10” to index the vegetable storage adequacy in China, while the other dataset sets the range “100-500” to index the animal product storage adequacy the same place.
 - b. Through processing the data and standardizing the value scale of various dataset, we can have a more clear background data for the further coding and analyzing, while avoiding some unnecessary and complex result which is hard for the final analysis.
4. **Coding to satisfy our challenge goals and aims.** Based on the challenge goals and targets we want to achieve through this final project, this step helps us to meet the purposes, while also gives us plenty of free rein to apply our knowledge gained during CSE 163 into our project.
- 5.
6. Currently, we will get an index by multiplying the prior two values and dividing by the later one. Since there are inherent differences in the weight of the food. We will process the data by dividing them by the median.(Example: (MedianFruit: 10 Median Vegetable: 20, MedianMeat:30) The country's data: Fruit: 20, Vegetable 20, Meat: 30. The country's index will be $2*1/1=2$). We decide to label each country by tags such as “vegan-like”. Then, we will generally classify the region's climate and religion. We identify religion into two kinds: special diet habits required religion and no diet habits required or no religion. They will correspond to a positive constant, and 0. They will be used to process the index, but not sure how accurate it will be. The design of a specific algorithm will come later. We also want to include climates into our database. But currently we only have hypotheses about how this factor works together with diet habits to influence public health. Lastly, we will use a regressor to make the analysis. We want to use some libraries which enable us to generate a dynamic map. The final map is supposed to be this: users provide a number, the program will generate a map of the health index of different regions at the given year and the predicted proportion of different kinds of diseases in the total population of the given year. When

moving mouse onto a region the historical change of the index and the historical change of promotion of different kinds of diseases will be presented at the right side of the map with a line chart.

“”””

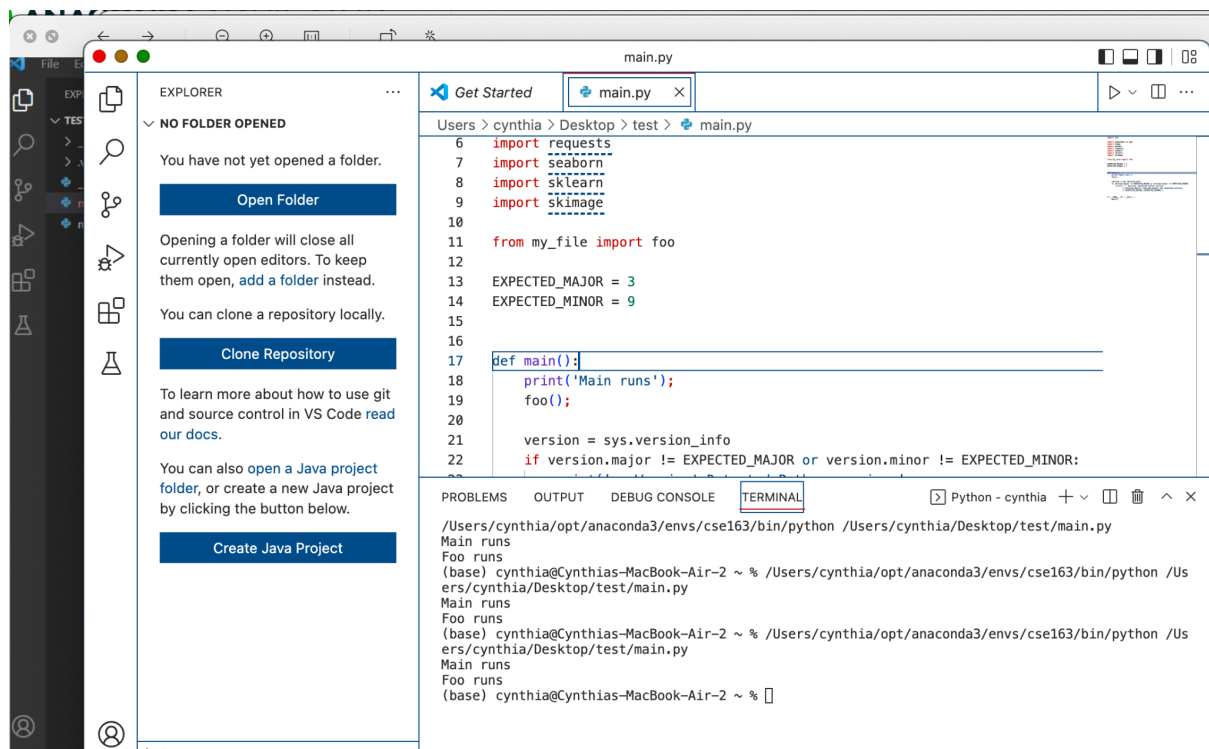
Work Plan:

There are mainly four parts of our experiment:

1. Setting up all materials. Following the proposal and discussion result with the TAs, we will try to prepare all necessary settings and sources. Each of us will download a VSCode on our computer, download a git, and create a repository on UW git lab. Then we will merge all necessary data into one or two CSV files. No more than three for efficiency. We also will write programs and collect data by ourselves if the wanted data cannot be found online. (Time: 4h)
2. Process data: We will check the raw data we get in the first step, and manually fix some problems. Then, using the files and python to process raw data. The processed data will turn several values in the raw data into one indicative variable (those indexes) according to the designed algorithms. (Time: 8h)
3. Coding, Visualization, and Training: Using libraries, we will make our programs which can do data visualization and machine learning work. (Time: 12h)
4. Testing: We will use different data sets to test our program's stability, accuracy, and efficiency. (Time 4h)

Libraries may be used: Seaborn, Panda, Scikit-Learn, geopanda, BeautifulSoup...

Proof of Setup



The screenshot shows the Visual Studio Code (VS Code) interface. On the left is the Explorer sidebar with options like 'Open Folder', 'Clone Repository', and 'Create Java Project'. The main editor area displays a file named 'main.py' with the following Python code:

```
6 import requests
7 import seaborn
8 import sklearn
9 import skimage
10
11 from my_file import foo
12
13 EXPECTED_MAJOR = 3
14 EXPECTED_MINOR = 9
15
16
17 def main():
18     print('Main runs');
19     foo();
20
21     version = sys.version_info
22     if version.major != EXPECTED_MAJOR or version.minor != EXPECTED_MINOR:
```

Below the editor is a panel with tabs for 'PROBLEMS', 'OUTPUT', 'DEBUG CONSOLE', and 'TERMINAL'. The 'TERMINAL' tab is active, showing the output of running the script:

```
/Users/cynthia/opt/anaconda3/envs/cse163/bin/python /Users/cynthia/Desktop/test/main.py
Main runs
Foo runs
(base) cynthia@Cynthias-MacBook-Air-2 ~ % /Users/cynthia/opt/anaconda3/envs/cse163/bin/python /Us
ers/cynthia/Desktop/test/main.py
Main runs
Foo runs
(base) cynthia@Cynthias-MacBook-Air-2 ~ % /Users/cynthia/opt/anaconda3/envs/cse163/bin/python /Us
ers/cynthia/Desktop/test/main.py
Main runs
Foo runs
(base) cynthia@Cynthias-MacBook-Air-2 ~ %
```

The screenshot shows the Visual Studio Code editor with a file named `main.py` open. The file contains the following code:

```
1 import sys
2
3 import geopandas
4 import numpy
5 import pandas
6 import requests
7 import seaborn
8 import sklearn
9 import skimage
10
11 from my_file import foo
12
13 EXPECTED_MAJOR = 3
14 EXPECTED_MINOR = 9
15
16
17 def main():
18     print('Main runs')
19     foo()
20
21     version = sys.version_info
22     if version.major != EXPECTED_MAJOR or version.minor != EXPECTED_MINOR:
23         print('Warning! Detected Python version '
24               f'({version.major},{version.minor}) but expected version '
25               f'({EXPECTED_MAJOR},{EXPECTED_MINOR})')
26
27
28 if __name__ == '__main__':
29     main()
30
```

The terminal output shows the following commands and results:

```
Microsoft Windows [版本 10.0.19044.1645]
(c) Microsoft Corporation. 保留所有权利。

D:\CSE163 Final Project\test>D:/anaconda/Scripts/activate

(base) D:\CSE163 Final Project\test>conda activate cse163

(cse163) D:\CSE163 Final Project\test>D:/anaconda/envs/cse163/python.exe "d:/CSE163 Final Project/test/main.py"
Main runs
Foo runs

(cse163) D:\CSE163 Final Project\test>
```

The screenshot shows the same Visual Studio Code editor with the `main.py` file. The code is identical to the previous screenshot. The terminal output shows multiple executions of the script:

```
PS C:\Users\TODD\Desktop\test> & C:/Users/TODD/anaconda3/envs/cse163/python.exe c:/Users/TODD/Desktop/test/main.py
Main runs
Foo runs
PS C:\Users\TODD\Desktop\test> & C:/Users/TODD/anaconda3/envs/cse163/python.exe c:/Users/TODD/Desktop/test/main.py
Main runs
Foo runs
PS C:\Users\TODD\Desktop\test> & C:/Users/TODD/anaconda3/envs/cse163/python.exe c:/Users/TODD/Desktop/test/main.py
Main runs
Foo runs
PS C:\Users\TODD\Desktop\test> & C:/Users/TODD/anaconda3/envs/cse163/python.exe c:/Users/TODD/Desktop/test/main.py
Main runs
Foo runs
PS C:\Users\TODD\Desktop\test>
```