**Group\_M\_Intakes**

**Names of all project participants (including emails)**

**Group member:**

Xinyao Huang [xh2612@columbia.edu](mailto:xh2612@columbia.edu)

Jiayi Song [js6367@columbia.edu](mailto:js6367@columbia.edu)

Juanyu Liu [jl6576@columbia.edu](mailto:jl6576@columbia.edu)

Rou Xie [rx2228@columbia.edu](mailto:rx2228@columbia.edu)

**Brief description of the proposed visualizations / analysis**

**Abstract:**

Our project aims to uncover potential links between the two by analyzing global, national, and New York City data, exploring and visualizing the relationship between happiness and per capita sugar intake. To quantify happiness, we used the Twitter Mood dataset for mood analysis to judge individual and group happiness. By collecting and collating statistics on sugar intake, we aim to build a visual map showing the relationship between happiness and sugar intake in different geographical locations.

**Link to data sources / API etc.**

**Dataset description:**

1. **Supply Utilization Accounts Food and Diet 2010-2020**

URL: <https://www.fao.org/faostat/en/#data/SUA>

The dataset presents average daily per capita macro and micro nutrient availability by country and by FAO/WHO GIFT food groups, expressed in grams/milligrams/micrograms/kcal per capita per day. It covers the majority of countries around the world.

1. **Household Consumption and Expenditure Surveys**

URL: <https://www.fao.org/faostat/en/#data/HCES>

The food and nutrient apparent consumption statistics were computed by the FAO Statistics Team from Household Consumption and Expenditure Survey (HCES) data. It mainly covers developing countries in South America, Africa and Middle West Asia.

1. **National Health and Nutrition Examination Survey (NHANES) 2015-2016**

URL: <https://www.globaldietarydatabase.org/management/microdata-surveys/718>

NHANES is designed to assess the health and nutritional status of adults and children in the United States. It collects data on a wide range of health-related topics, including diet, physical activity, obesity, chronic diseases, and environmental exposures.

# **Food Consumption and Nutrient Intakes 2015-2018**

URL: <https://www.ers.usda.gov/data-products/food-consumption-and-nutrient-intakes/>

ERS of USDA provides data on food consumption and nutrient intake by food source and demographic characteristics.

1. **Sentimental score:** 
   1. **Twitter Sentiment Geographical Index**

<https://www.globalsentiment.mit.edu/dataset>

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/3IL00Q>

**Data structure:**

DATE---- date ---- the date of the sentiment index

NAME\_0 ---- string ---- the country name

NAME\_1 ---- string ---- the state/province name

NAME\_2 ---- string ---- the county/city name

SCORE ---- float ---- a float value between 0 and 1 representing the sentiment index where 1 represents a positive sentiment and 0 represents the negative sentiment.

N ---- int ---- the number of posts generated given the specific date

The raw tweet data used to produce the global sentiment and geography index dataset (GSGD) is from Harvard CGA Geotweet Archive v2.0, a global collection of geotagged tweets spanning time, geography, and language maintained by the Harvard Center for Geographic Analysis. The Archive extends from 2010 to the present and is updated daily. The number of tweets in the collection is approximately 10 billion.

**Pros/cons:**

Precalculated sentiment score.

Calculated by cites.

* 1. **World Happiness Report**

<https://worldhappiness.report/data/>

**Data structure:**

The Gallup World Poll questionnairemeasures 14 areas within its core questions: (1) business & economic, (2) citizen engagement, (3) communications & technology, (4) diversity (social issues), (5) education & families, (6) emotions (well-being), (7) environment & energy, (8) food & shelter, (9) government and politics, (10) law & order (safety), (11) health, (12) religion & ethics, (13) transportation, and (14) work.

**Pros/cons:**

Precalculated sentiment score.

More variables.

Calculated by countries

**Technique:**

For this project, a combination of statistical analysis, natural language processing (NLP), and data visualization techniques will be employed. We plan to use NLP to analyze the sentiment of tweets from the Twitter Sentiment Geographical Index, extracting happiness levels across different geographical locations. Techniques such as sentiment analysis and topic modeling may be applied to categorize tweets and understand the context of happiness and sugar intake discussions. Statistical methods will be utilized to correlate happiness levels with per capita sugar intake, employing regression analysis or correlation coefficients to identify significant relationships. Visualization techniques will include the use of *ggplot2* for generating line charts to illustrate trends over time, and interactive mapping tools to create heat maps and interactive maps that geographically display the relationship between happiness and sugar intake across various regions. Data manipulation and cleaning will be facilitated by tools like *dplyr* n R or *pandas* in Python, ensuring the datasets are ready for analysis.

**Visualization type:**

The project aims to employ a range of visualization types to effectively communicate the findings:

* **Line Charts:** To depict trends over time in both happiness and sugar intake across different countries and regions, allowing for temporal comparisons.
* **Interactive Maps:** To visually represent the geographical distribution of happiness and sugar intake levels, enabling users to explore data across different locations interactively.
* **Heat Maps:** To illustrate concentrations of high or low happiness levels in correlation with sugar intake, providing a clear visual indicator of any potential hotspots.
* **Scatter Plots**: To analyze the relationship between happiness and sugar intake, potentially with trend lines to indicate the direction and strength of the relationship.
* **Bar Charts:** For comparing happiness and sugar intake levels across different demographic segments or countries.

**Potential problems/ limitations**

* **Data Quality and Availability**: The reliability of self-reported data on sugar intake and the sentiment scores derived from social media may present challenges. Variations in data quality across different sources and potential biases in self-reporting or sentiment analysis algorithms could affect the accuracy of the findings.
* **Cultural and Linguistic Differences:** The interpretation of happiness and discussions around sugar intake may vary significantly across cultures, which could influence sentiment analysis results. Language differences and the use of slang or idiomatic expressions on social media platforms like Twitter may also pose challenges for accurate sentiment analysis.
* **Correlation vs. Causation:** Establishing a correlation between happiness and sugar intake does not imply causation. There may be underlying factors influencing both variables that are not accounted for in the analysis.
* **Ethical and Privacy Concerns:** The use of publicly available social media data raises ethical considerations around privacy and consent, especially when analyzing and presenting data at the individual or community level.