Nutritional Profile of Countries Around the World

The Nourishing Four ʕ •ᴥ• ʔʕ •ᴥ• ʔʕ •ᴥ• ʔʕ •ᴥ• ʔ

Sahil Bansal (sbansa12), Vanessa Chang (vchang5), Christine Han (chan14), Liza Kolev (lkolev)

Hypothesis/Goal/Introduction

Food is a vital part of one's survival. We set out to investigate how the nutritional profiles of countries around the world differed from each other and what factors may influence a countries' nutritional profile. We looked into how GDP per capita (purchasing power parity 2017) (henceforth GDP) affected fat percentage in diets, how alcohol consumption and life expectancy correlated to each other, and how GDP and meat consumption were correlated to each other. In addition, we clustered countries based on their nutritional profile and, given GDP, predicted the percentage of fat in a countries' diet.

Data

We obtained our data from two main sources: Our World in Data, and Food and Agriculture Organization of the United Nations (FAO). It is important to note that the data from Our World in Data traces back to FAO. The sourced information about the caloric composition of each countries' diet, each countries' GDP, life expectancy, and import and export data for a range of years. Then, we put all the data in one database by joining on countries and years. Afterwards, we identified and removed unused columns. Finally, we limited the range of years from 2000 to 2013 as we were focusing on the 21st century and data from 2013 onwards was sporadic.

Results and Analysis

Claim #1: A country whose GDP is above $22,855 will have a higher fat percentage in their diet compared to a country whose GDP is below $22,855.

Support for Claim #1: We first categorized the countries into above and below the threshold, and then ran a **two sample t-test**. We ran the test every year from 2000 to 2013 and obtained p-values close to 0. Given that the p-values of the tests were less than 0.05, we reject the null hypothesis i.e. we reject that there is no significant difference in the mean percentage of fat in the diet between the two countries based on the GDP per capita. In other words, this means that the difference in the mean percentage of fat in the diet between the two groups is unlikely to have occurred by chance alone and is likely due to the difference in GDP. That being said, this is only speaking to statistical significance and not necessarily practical significance.

Claim #2: There is a negative correlation between alcohol consumption and life expectancy.

Support for Claim #2: We chose to use the **Spearman rank-order correlation** to calculate the p-value of our hypothesis and the correlation between the data. Generally, Spearman correlation measures the monotonicity of the two datasets, and we validated the monotonicity of the data via our scatter scripts. In addition, we used the **one sample t-test** to evaluate the correlation because the population parameters of our data are unknown and the data being analyzed is not categorical. While some countries' life expectancies have a negative correlation with their daily alcohol consumption, this is not the case for all countries. Many factors account for this, but the main factor would probably be that alcohol consumption is not the only cause of shortened life expectancy. In fact, some cultures believe drinking a glass of wine a day lengthens life expectancy due to the antioxidants within them. Additionally, the similarity of results from correlation across countries per year and the correlation across countries and year suggests that, overall, there is a **positive** correlation between daily alcohol consumption and life expectancy.

Claim #3: There is a positive correlation between GDP and the amount of meat consumed in a country.

Support for Claim #3: We chose to use the **Spearman rank-order correlation** to calculate the p-value of our hypothesis and the correlation between the data. Generally, Spearman correlation measures the monotonicity of the two datasets, and we validated the monotonicity of the data via our scatter scripts. In addition, we used the **one sample t-test** to evaluate the correlation because the population parameters of our data are unknown and the data being analyzed is not categorical. Though the correlation across years for each country fluctuates a lot, what is the most relevant to us for determining whether or not to reject the null hypothesis is the correlation across countries and years. From this, we can see the correlation to be about 0.88, or positive, which indicates that there is a **positive** correlation between GDP and the amount of meat consumed in the country.

Model + Evaluation Setup

Model 1: We want to build an ML model that, given a country's GDP per capita purchasing power parity 2017, would be able to predict the percentage of fat in that country's diet. We used numpy's **polyfit** function, with varying degrees, to find the best regression line.The metric we used to measure the success or failure of the model was the **R-squared (R2) value**. We discovered from the results that using polyfit with degree 2 was the best fit for our data. This is because when we compare the R2 value from the different degree models, degree 2 was the minimum degree after which the R2 value did not significantly increase. Degree 2 had an R2 value of 0.54, which means that 54% of the variation in the dependent variable, percent of fat, is explained by the independent variable, GDP, in the regression model. This is the case both with and without validation.

Model 2: We wanted to cluster countries based on the main attributes of a nutritional profile in order to figure out which countries had similar diets. These clusters could then be used in future for something like classification. For this model, we chose to use **KMeans** to cluster the data. Note that since we want to cluster countries, we produced cluster graphs for each year. So, based on the results of the KMeans clustering, we can see the countries that are close together based on the 3 attributes that we chose to cluster by. When clusters are compared across years, it seems like generally the clusters do not change, but a couple of countries do jump about from cluster to cluster which may be due to war, particularly bad weather, economic crisis and many other factors that are not accounted for.