|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
|  |  |  |  |
| Global food prices  impact analysis on population, mORTALITY, bIRTH- rATE and GDP  Group Assignment – Data Science 2 – Introduction to Statistics | |
|  | Presented by - Group 2  Jobbin Samuel  Ricky Lam  Maxim Smetin | Ramila Mudarth  Avery Lizhong Wang  James Lising |  |

Table of Contents

[Objectives 3](#_Toc90209416)

[iNTRODUCTION 3](#_Toc90209417)

[Data preparation 3](#_Toc90209418)

[data ANALYSIS 5](#_Toc90209419)

[Linear Regression Analysis: Rice and Birth Rate 6](#_Toc90209420)

[Casual INference: Asia 7](#_Toc90209421)

[Birth rate and Rice Price 7](#_Toc90209422)

[ASIA: Child Mortality rate and Rice Price 8](#_Toc90209423)

[Multi-variable Linear regression Analysis 8](#_Toc90209424)

[Birth Rate analysis 8](#_Toc90209425)

[Child Mortality Analysis 9](#_Toc90209426)

[DEATH Rate Analysis 10](#_Toc90209427)

[CONCLUSION 11](#_Toc90209428)

[Bibliography 13](#_Toc90209429)

[ApPENDIX 14](#_Toc90209430)

# Objectives

In this project we are going to explore a question about possible impact of food prices into population. Is there correlation between changes in food prices and birth rate, death rate and child mortality?

We choose dataset with food prices for the last twenty years in developing countries, as well as birth rate, death rate and child mortality. Moreover, how GDP impacts population growth in developing countries?

# iNTRODUCTION

The global food prices have always been subjected to external influences like fuel prices, natural disasters as a result of global warming activities. Some countries have been affected more than others depending on their ability to endure the fluctuations and food availability. There are assumptions made that the food prices influence the mortality, the GDP, affordability and income which in turn effects the population trends. Producers benefit from rise in prices where are consumers benefits from lower food prices. Any fluctuation in prices will have an effect especially on the lower income individuals as a result causing food shortages.

The objective of our assignment was to determine the impact of food prices of specific commodity in developing countries and compare their GDP, Mortality Rate to determine any kind of correlation that might exists using last twenty years of data. Additionally, we intend to pair it with restaurant index data to see if there are any insights derived.

We intend to resolve the two hypotheses

* **Hypothesis 1:**How much food price influence Population? Null Hypothesis is food price isn’t a key driver of population. The alternative Hypothesis is that food price somewhat affect population.
* **Hypothesis 2:**How much do food prices impact all parameters? Null Hypothesis is that food prices impact all parameters equally. The alternative hypothesis is that there are some differences between some parameters affected by food prices.

# Data preparation

Preparation of the data set required compiling and sourcing from multiple location. Each data set had to be solved for challenges presented and transformed for the required analysis

A picture containing text, businesscard, screenshot

Description automatically generated

Figure 1 Data Schema

|  |  |  |
| --- | --- | --- |
| The core dataset contains Global Food Prices data from the World Food Programmed covering foods such as maize, rice, beans, fish, and sugar for 98 developing countries and some 1,500 markets. The data goes back as far as 1992 for a few countries, although many countries started reporting from 2003 or thereafter. The Data is collected by [WFP(The World Food Program)](https://www.wfp.org/) and the dataset was distributed by [HDX](https://data.humdata.org/) . Data includes developing countries, locality, market, goods purchased, price & currency used, quantity exchanged, and month/year of purchase.  The unit of measure (UOM) for the commodities were not consistent per observation. We have created a table for normalizing the retail value to kilograms. This step let us to perform aggregation of food prices across the whole dataset per Continent, Country, Commodity and Year. A conversion table was manually created to do the math. No conversion was applied to fuel, and commodity item such as toothbrush, toothpaste, and other liquid that is unmeasurable. UOM in file has not been changed in the file. But those that have been change will be reflect in retail price. *e.g. where the retail was 2500 dollar for 12 kg was converted to 208 dollars per KG.*    Figure 2 The data frame structure  The Food prices were in local currency. The classification of the commodity category was too granular, so we had to aggregate it at a boarder level*. e.g. Rice commodity name has 82 different versions. But we created a category by using the first word in the string.*  The other features merged into the table are   * Regional Classification - Countries were reclassified by the region for higher level analysis and to obtain greater number of observations eg. Northern Africa, Eastern Africa, Middle Africa, Southern Africa. * Gross domestic product (GDP per capita) – Twenty years of GDP information trend per country * Child Mortality Rate - Child Mortality counts since 1967 per 1,000 * Birth and Death Rate – Birth and Death count per country since 1960 per 1,000 * Fertility Rate - Fertility Rate since 1950, shown per woman   We added another layer of classification for the commodities   * Raw – Milk, Eggs, Rice * Processed – Bread, Curds * Other - Fuel, Internet  data ANALYSIS By splitting to categorical view, there is no relationship between commodity, & price with any value/ table shown in the matrix below    Figure 3 - Correlation Matrix for Categorical Variables  The prepared data set was further compressed and aggregated to give commodity per country per area. In the prepared data set, we observed that Rice has been represented in most of the countries. Therefore, we used rice prices as they reflect food prices across the dataset. We created a subset data frame to run an analysis on. A heat Map was generated using this data set.    Figure 4 Evaluation of Pearson Coefficient across the data frame  Initial high-level observations from the Heat Map shows positive correlations between the following parameters   * Birth Rate, Fertility, Death Rate and Child Mortality – suggesting that they are colinear. This is a natural observation, for example, we would expect birth rate to depend on fertility.   The negative correlation exists in the following parameters   * GDP per capita and birth rate, child mortality, fertility, death rate * Price of Rice versus child mortality, birth, and death.   We are going to have deeper dive into relation between rice price and birth rate. Rice being chosen as it is present across our dataset. This would mean that this is a common meal ingredient across most of the countries, and its price fluctuations is colinear with other grocery ingredients. Therefore, the rice price was taken as a representative of the global food prices. Here we want to study how the food price impacts population growth, namely birth rate. Linear Regression Analysis: Rice and Birth Rate Once our dataset was built, we can check for relation between rice price and birth rate Figure 5. Overall data is quite scattered, and if we run linear regression, we most likely will not see significant relation between the two parameters. Indeed, heat map (Figure 4), shows Pearson coefficient between them is -0.071, and slope has high p-value – we failed to reject null hypothesis – rice price doesn’t impact birth rate. Our dataset represents developing countries – they are quite different. What if we will have a closer look at this dataset and zoom into Asia continent.    Figure 5 Scatter plot with rice cost and birth rate distribution across the dataset. Casual INference: AsiaBirth rate and Rice Price Whole Asian part of the dataset have birth rate and rice price quite disperse distribution on both scales. For example Figure 6, shows the same data but with different granularity level – starting from continent Asia (Figure 6A), by regions (Figure 6B), by countries (Figure 6C) and one country – Afghanistan (Figure 6d). We observed that the correlation between two parameters become stronger if regions and countries being includes.  Linear regression analysis shows no relationship between birth rate and rice on the continent level. Pearson squared coefficient is equal to 0.07, p-value for rice coefficient is 20%. We fail to reject null hypothesis – there are no influence of rice price on the continent level.  Meanwhile, when we examine our relationship on region level (Figure 6B) – the intercept and slope values for each region are not the same. In case of Central Asia – the birth rate is independent of rice price.    Figure 6. Average rice price and birth rate in Asia, Asia regions and Asian countries    Once we increase further the granularity and have the distribution by countries – the diversity becomes obvious. Each country will have different intercept – mean values. However, the slope has negative sign – suggesting that with increase of the rice will cause decline in the birth rate.  For example, Afghanistan (Figure 6D) R squared is equal to 0.62, the coefficient is -45.2 and p-value 1%. Therefore, with 99% confidence we reject null hypothesis and conclude that there is an impact of the food price on the birth rate in Afghanistan. ASIA: Child Mortality rate and Rice Price Let’s examine if there are relationship between food prices and child mortality. On the continent level the data has a strong spread (Figure 7A), and no correlation (R2 = 0.115). Rice slope coefficient has negative 0.95 value. But, even from this level, one can see that in case of small child mortality – between 0 and 2 per thousand – the rice price has no impact. However, as the child mortality becomes more then 2 the rice role becomes stronger.  From region level we can see that rice price impacting Southern Asia countries (Figure 7B). Figure 7C shows the relationship on the country level. For some countries the data points are almost perpendicular to x-axis – showing indeed no dependency from y-axis (rice price). We were happy to see that, as this mean that food accessibility doesn’t contribute to the child mortality in those countries. However, in case of Afghanistan R-squared is equal to 0.62, and slope for rice price defined with 1.0% p-value. In this case child mortality can be expressed:    Figure 7 Average rice price and child mortality in Asia, Asia regions and Asian countries Multi-variable Linear regression Analysis We performed multivariable Linear regression modeling on Birth Rate, Death Rate and Mortality at a regional level to see the relationship between the parameters in the data Birth Rate analysis In the Birth Rate OLS Model – A , we observed adj R Squared is 68% making the model reasonable fit.  All the p values are zero as a result we can say, we accept the null hypothesis. All parameters are acceptable in the model.    Figure 8 Birth Rate analysis  We ran Birth Rate OLS Model – B, after removing the outliers from the Influential Plot and the model had an Adjusted R Square of 72.3%.    Figure 9 Influential plot Child Mortality Analysis In the Child Mortality analysis, the Adjusted R-Squared is 15% as a result the null hypothesis is unacceptable.    Figure 10 Child mortality analysis DEATH Rate Analysis The model is at 52%. The price of Rice is not relevant as we need to remove the parameter as the p value is greater than 0.05    Figure 11 Death rate analysis CONCLUSION *What did we infer from our analysis?*  We have studied relations between food prices and different aspects of population changes – birth rate, death rate and child mortality. We explored GDP contribution as well.  Having chosen rice as the most representative commodity we have inferred what it has negative correlation with population.  Overall dataset has small correlation between food price and birth rates. However, if we introduce continents and countries, and apply inferred analysis there, we have observed strong correlation between food price and birth rates.  We have observed that the birth rate in Afghanistan, can be represented through linear regression in the following expression:  inside our dataframe, and therefore reflecting   * **Hypothesis 1:**Is food price influence Population? Null Hypothesis is food price isn’t a key driver of population. The alternative Hypothesis is that food price does affect population, namely birth rate.   We are rejecting null hypothesis 1. The relationship between food price and birth rate can be expressed with linear regression on the country level. P-values for slope coefficients are less then 1%.   * **Hypothesis 2:**How much do food prices impact all parameters? Null Hypothesis is that food prices impact all parameters equally. The alternative hypothesis is that there are some differences between some parameters affected by food prices.   We failed to reject null hypothesis 2, since there one of three population parameters – child mortality did not have **correlate with rice price.** |  |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
|  |

# Bibliography

*Exchange Rates*. (n.d.). Retrieved from https://data.oecd.org/conversion/exchange-rates.htm.

*Global Food Prices*. (n.d.). Retrieved from https://data.humdata.org/dataset/wfp-food-prices.

*Gross Domestic Product*. (n.d.). Retrieved from https://stats.oecd.org/Index.aspx?DatasetCode=SNA\_TABLE1.

*Population Trends*. (n.d.). Retrieved from https://data.worldbank.org/indicator/SP.POP.TOTL.

Ritchie, M. R. ((2013) ). *Food Prices". Published online at OurWorldInData.org*. Retrieved from 'https://ourworldindata.org/food-prices' [Online Resource].

# ApPENDIX