How does the cost of different food types affect energy-sufficient meals in various countries?

for the Bachelor of Science Honours Degree in
Financial Mathematics and Industrial Statistics

By
E.G.M.Lakruwan
SC/2020/11795

Supervisor:
Prof. Leslie Jayasekara

Department of Mathematics
University of Ruhuna
Matara.

2023

DECLARATION

I, E.G.M.Lakruwan, declare that the presented project report titled, ”How does the cost of different food types affect energy-sufficient meals in various countries?” is uniquely prepared by me based on the group project carried out under the supervision of Prof.Leslie Jayasekara Department of Mathematics, Faculty of Science, University of Ruhuna, as a partial fulfillment of the requirements of the level II Case Study course unit (MIS2231) of the Bachelor of Science Honours Degree in Financial Mathematics and Industrial Statistics in Department of Mathematics, Faculty of Science, University of Ruhuna, Sri Lanka. It has not been submitted to any other institution or study program by me for any other purpose.

Signature:. . . . .
Date:. . . . .

SUPERVISOR’S RECOMMENDATION

I/We certify that this study was carried out by E.G.M.Lakruwan under my/our supervision.

Signature:. . . . .
Date:. . . . .
Prof.Leslie Jayasekara,
Department of Mathematics,
Faculty of Science,
University of Ruhuna.

ACKNOWLEDGMENT

First of all, I would like to thank all of those who helped me in completing this report. Firstly, I would like to express my sincere gratitude to my supervisor, Prof. Leslie Jayasekara, for guiding me throughout this project and supporting me in finding and fixing my mistakes. Also, I'd like to thank our course coordinator, Dr. Lakshika Jayathilake, and all the other lecturers who helped me every step of the way to make this work a success. They offered their advice and help without hesitation whenever I needed it. Their expertise and advice were invaluable, and I am truly grateful to have such great mentors.

I would also like to thank my team members for all of their help and hard work. We were able to learn a lot together and help each other out. Also, all the friends who helped me in various ways, I am truly grateful that I had all their help which helped the succession of this report. I could not have completed it without you.

Thank you.

Contents

List of Figures

1 Introduction

1.1 Background of the Study

A meal is what keeps a man alive. What people eat affects directly to their health. Also, some professions require specific meals, and some lifestyles ask for different nutrition. Most of the studies out there are on healthy food and healthy diet plans. But having enough energy is as important as taking a healthy meal. Thus, we have decided to study the cost and affordability of an energy-sufficient diet.

In a high-income country, it might be easier to afford a healthy meal. But a low-income country might not be as fortunate as them. A person can live on a not-so-healthy meal if they get to eat enough to survive, but a person needs enough calories in order to function properly. A meal where we get enough calories to properly function is an energy-sufficient diet. According to the NHS (https://www.nhs.uk/) many things decide our calorie intake. “An ideal daily intake of calories varies depending on age, metabolism, and levels of physical activity, among other things.” But generally, it is “between 2000 to 2500 calories per day”. And there are many ways we can get those calories. The World Data Bank has collected data on how much money it will cost for such a meal (an energy-sufficient diet). This includes the cost of the full meal, and the cost of the food types such as vegetables and fruits cost per meal, in many countries. We used these data for the study of how the cost of different food types affects energy-sufficient meals in various countries.

1.2 Why is this Important

We know that there are weaknesses in energy-sufficient diet plans all over the world, as well as countries that cannot afford an energy-sufficient diet. If we can find a way to provide enough calories to those communities for an affordable price, it will be very productive and helpful for each country to develop. If we can find what foods have the most impact on the cost of the meal, the governments can take suitable actions depending on the findings. It will help to keep a stable economy while providing the country folk with enough calories.

Additionally, this research contributes to the broader discussions on global economic disparities and social equity. It highlights the disparities in access to nutritious food based on income levels and geographical regions. By identifying the factors that cause these disparities, the study provides a foundation to make decisions on reducing food inequalities.

1.3 Objectives

In our study we wish to find answers for the following questions.

- What are the food types which affect most the cost of an energy-sufficient diet?
- What is the significant impact of the cost of various food types on an energy-sufficient diet?

By using the data collected from the World Data Bank, we wish to study the relationship between the cost of each food type in a meal and the cost of the meal as a whole. Also, to analyze the multiple linear regression of them on a global scale.

2 Literature Review

One of the research problems that needs to be solved is how to model the cost of energysufficient diet behavior. It is necessary to study energy-sufficient determinants because of their important role in the world of energetic persons, the productiveness of a person in human work, and world development.
The research article published by Melissa Bateson, Clare Andrews, Jonathan Dunn, and Daniel Nettle who are the group of authors in 2021 says that energetic requirements in foods may depend on the body mass of a person, the age of the person, and the gender he/she belongs to. Furthermore, the energy consumption level of a person may depend on their lifestyle. For example, if we consider a sportsman and a software engineer, sportsmen need more energetic meals than software engineers. Because a sportsman wastes his calories more than a software engineer.
Another research article published by Patrik Webb, Goodarz Daniel, William A. Masters, Katherine L. Rosettie, and Sarah Kranz who are the group of authors in 2019 shows that the cost of energy-sufficient meals affects malnutrition. The countries that have low national income levels have a lack of people who eat energy-sufficient meals. Furthermore, they said that the difference between a healthy meal and an energy-sufficient meal is the proportion of nutrition parts that are the proteins and glucose levels. It describes that the glucose proportion and protein proportion of an energetic meal is the triple amount of the corresponding proportions of a nutritious meal. Because the vitamin and mineral proportions are effects on nutritious meals. Furthermore, a healthy meal is a balanced meal of nutritious parts and an energetic meal is an unbalanced meal of nutritious parts and the cost of an energetic meal is higher than the cost of a healthy meal.
In this study, we wish to further investigate into this matter on a scale of how the cost of each food type affects the cost and affordability of an energy-sufficient diet.

3 Materials and Methods

3.1 Population Sample

In the study, we chose the population as the countries in the world. We got the data from the World Data Bank, so one could state that we used convenient sampling to select the sample as it was what's convenient for us. But we can also argue that the data collection was done by the World Data Bank, in which case we do not know how they selected their sample. All the countries in the World Data Bank dataset were taken for the study.

3.2 Research Approach

This study uses a quantitative research approach to examine the relationship between food types' costs and the cost of an energy-sufficient meal in different countries. The approach is justified by research objectives, problem nature, data availability, and statistical analysis for more meaningful conclusions.

3.2.1 Research Objectives

The main goal of this study is to analyze how the cost of different food types affects the cost of an energy-sufficient meal globally using a quantitative research approach, focusing on numerical data analysis and statistical relationships between variables.

3.2.2 Nature of the Problem

The purpose of the study is to analyze cost variables. Both independent variables (cost of different food types) and dependent variables (cost of an energy-sufficient meal) among different countries. A quantitative method allows for efficient comparison, displays trends and patterns, and allows for effective comparison and investigation of energy-sufficient meal prices.

3.2.3 Data Availability

Quantitative data on the cost of different food types and the cost of energy-sufficient meals in various countries were available for analysis. This availability of numerical data makes a quantitative research approach the natural choice for this study.

3.2.4 Statistical Analysis Requirements

Given the need to assess the impact of multiple independent variables (cost of different food types) on a continuous dependent variable (cost of an energy-sufficient meal), multiple linear regression was identified as an appropriate statistical method for this research. Simple graphical representation has been used for exploratory data analysis.

3.3 Conceptual Framework

Independent variables are the cost of different food types. They are the cost of fruits, cost of vegetables, cost of starchy staples, cost of animal-source foods, cost of legumes, nuts and seeds, and cost of oils and fats. The dependent variable is the cost of an energy-sufficient diet as illustrated in Figure 1.

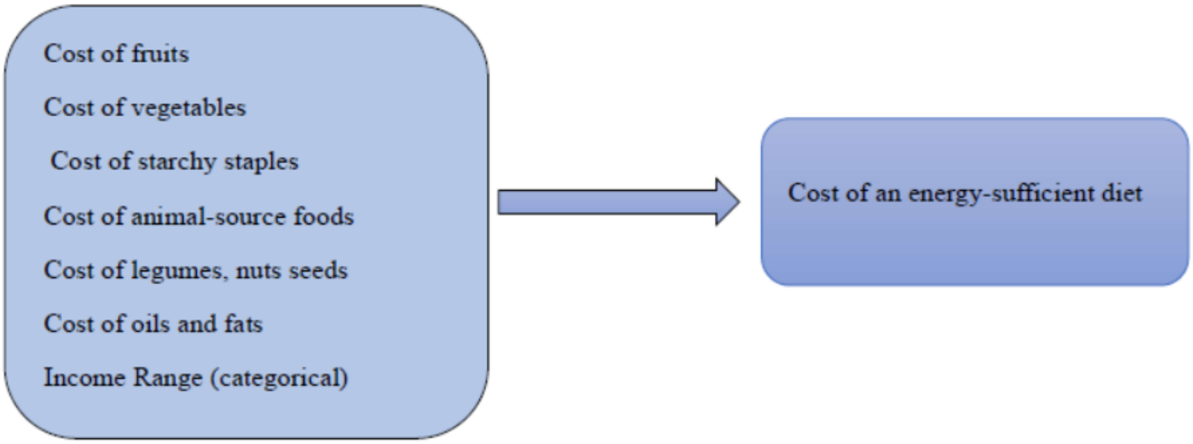

Cost of fruits
Cost of vegetables
Cost of starchy staples
Cost of animal-source foods
Cost of legumes, nuts seeds
Cost of oils and fats
Income Range (categorical)
Cost of an energy-sufficient diet

Figure 1: Conceptual Framework

3.4 Research Hypothesis

A research hypothesis is a statement of the relationship between two or more variables. It is a prediction of the outcome of a research based on the connection between the variables under consideration. The purpose of a research hypothesis is to guide a study and give a foundation for testing. Null hypotheses and alternative hypotheses are the two various types of research hypotheses. The null hypothesis (H_0) states that there is no relationship among the variables under consideration. Contradictory to that, the alternative hypothesis (H_a), there is a relationship between the variables.

According to our study,

H_0: There is no relationship between the costs of different food types and the cost of energy-sufficient meal.

H_a: There is a relationship between the costs of different food types and the cost of energy-sufficient meal.

4 Data

The relevant data was selected and downloaded from the World Data Bank's official website. Note that the cost of each item is valued in USD. Also, the cost of an energy-sufficient diet is measured to give sufficient calories indicated by the WHO, and the cost of each food type is determined for how much it will cost per each food type proportion that would be included in a healthy meal in WHO standards.

The categorical variable income range was measured using the GDP per capita. The ranges were selected according to standards provided by the World Bank Blogs.

- Low Income : $< \$1,045$
- Middle Income : $\$1,046 - \$12,695$
- High Income : $> \$12,695$

4.1 Data Cleaning Process

A total of 177 data was recovered originally. But when checking the data manually, we found that there were also continents included in the data. So, we had to manually remove them from the original dataset. Then using R-studio we cleaned the dataset to remove the missing values (code entered below) which reduced the sample to 143.

Out of 195 countries in the world (according to Google) we were only able to receive the data of around 170 countries. However, in the data cleaning process, we removed about 15% of the original data. But even after that, we had 143 data remaining, which is 74% of the population and 85% of the original sample. Since we are taking a sample for the data analysis, we decided that the cleaned data would be a sufficient sample for the study.

4.2 Dealing with Duplicates and Outliers

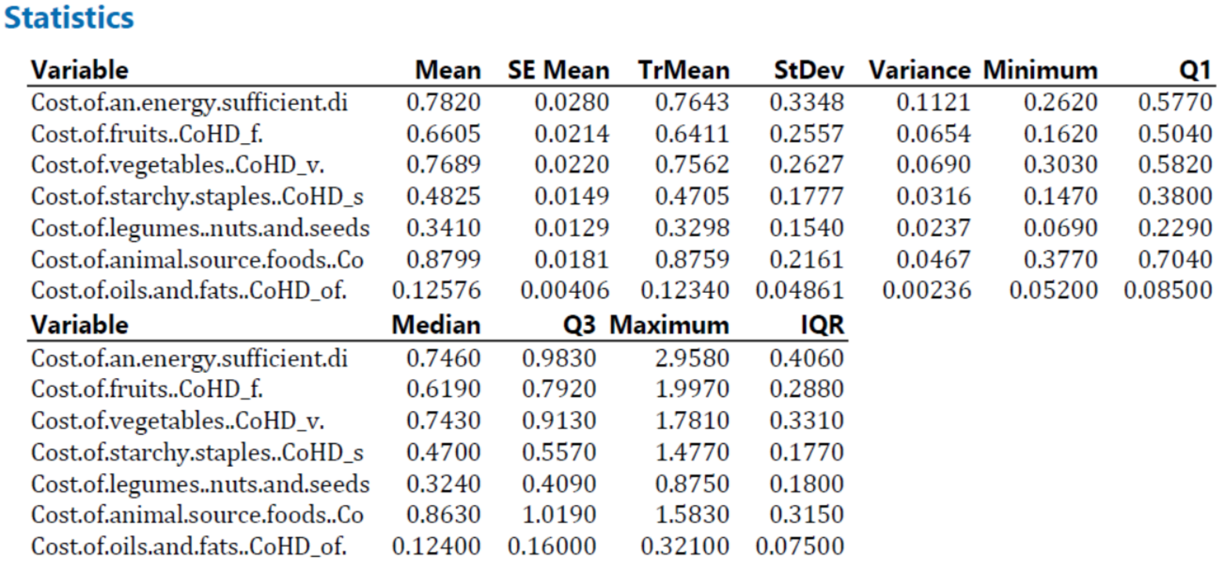
With regard to duplicate values, since we took data from the World Data Bank, and each data point represents a different country, we did not get any duplicates.

Then, before the descriptive analysis, we had to deal with the outliers. We used boxplots to identify the outliers in the data (Figure 19). There are many methods to deal with outliers., i.e. standardization, removing the outliers, etc. In the descriptive data analysis, we will further talk about how we dealt with the outliers.

5 Results

5.1 Exploratory Data Analysis

Now let's take a look at the data after the cleaning process in a graphical and statistical view.



Statistics

Variable	Mean	SE Mean	TrMean	StDev	Variance	Minimum	Q1
Cost.of.an.energy.sufficient.di	0.7820	0.0280	0.7643	0.3348	0.1121	0.2620	0.5770
Cost.of.fruits..CoHD_f.	0.6605	0.0214	0.6411	0.2557	0.0654	0.1620	0.5040
Cost.of.vegetables..CoHD_v.	0.7689	0.0220	0.7562	0.2627	0.0690	0.3030	0.5820
Cost.of.starchy.staples..CoHD_s	0.4825	0.0149	0.4705	0.1777	0.0316	0.1470	0.3800
Cost.of.legumes..nuts.and.seeds	0.3410	0.0129	0.3298	0.1540	0.0237	0.0690	0.2290
Cost.of.animal.source.foods..Co	0.8799	0.0181	0.8759	0.2161	0.0467	0.3770	0.7040
Cost.of.oils.and.fats..CoHD_of.	0.12576	0.00406	0.12340	0.04861	0.00236	0.05200	0.08500

Variable	Median	Q3	Maximum	IQR
Cost.of.an.energy.sufficient.di	0.7460	0.9830	2.9580	0.4060
Cost.of.fruits..CoHD_f.	0.6190	0.7920	1.9970	0.2880
Cost.of.vegetables..CoHD_v.	0.7430	0.9130	1.7810	0.3310
Cost.of.starchy.staples..CoHD_s	0.4700	0.5570	1.4770	0.1770
Cost.of.legumes..nuts.and.seeds	0.3240	0.4090	0.8750	0.1800
Cost.of.animal.source.foods..Co	0.8630	1.0190	1.5830	0.3150
Cost.of.oils.and.fats..CoHD_of.	0.12400	0.16000	0.32100	0.07500

Figure 2: Descriptive Analysis of Cleaned Data

The above statistics show the means, trimmed means, standard deviations, medians, inter-quartile ranges, and minimum-maximum values of the selected sample of 143 countries. From this output, we can get some idea of how the cost of an energy-sufficient diet and each food type have spread around the world.

For example, let's consider the cost of an energy-sufficient meal. The cost of such a meal varies from \$0.2620 to \$ 2.9580. But the IQR is from \$0.5770 to \$0.9830 which has a significant difference from the total range. Thus, we can imagine that the distribution of the cost is leftskewed. The mean and median being \$0.7820 and \$0.7460 respectively further support this ideal. This suggests that most of the countries cost \$0.5770-\$0.9830 per energy-sufficient meal, but there a a few countries that cost almost double, maybe triple the amount as well.

By all means, we can get a rough image of the cost of each food type throughout the world from this descriptive analysis.

Diving deep into the data, let's inspect each variable separately.

Cost of energy-sufficient food (dependet variable)

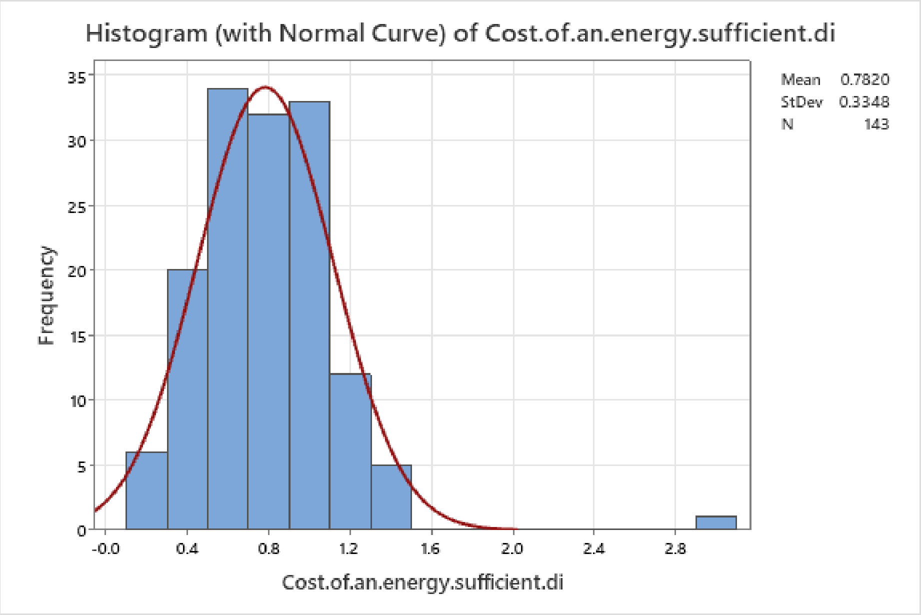


Figure 3: Histogram of Cost of an Energy-sufficient Diet

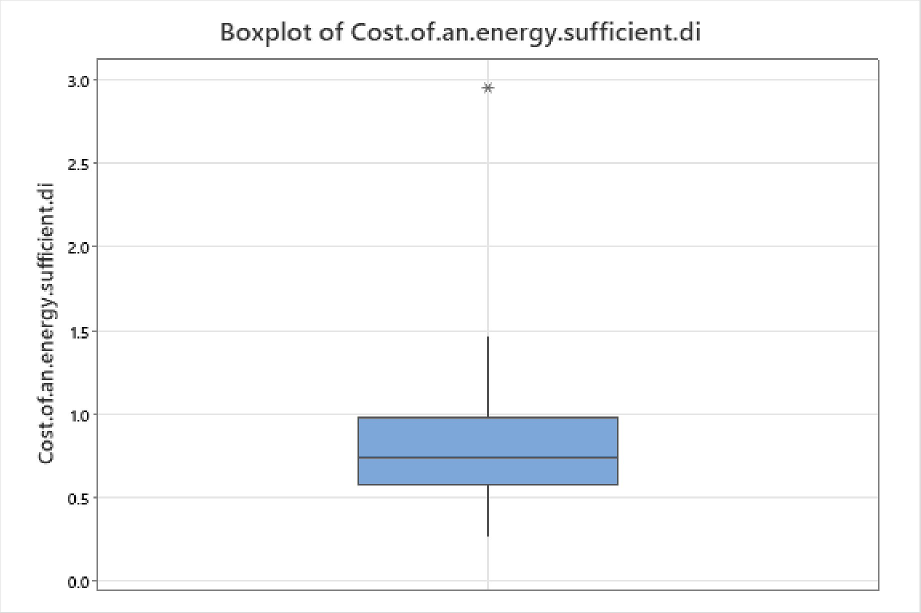


Figure 4: Boxplot of Cost of an Energy-sufficient Diet

The above figures suggest a left-skewed distribution for the selected data for the dependent variable. We can clearly see there's an outlier amongst the data. One can suggest that the data could've been normally distributed if it weren't for the outlier.

Cost of Fruits (independet variable)

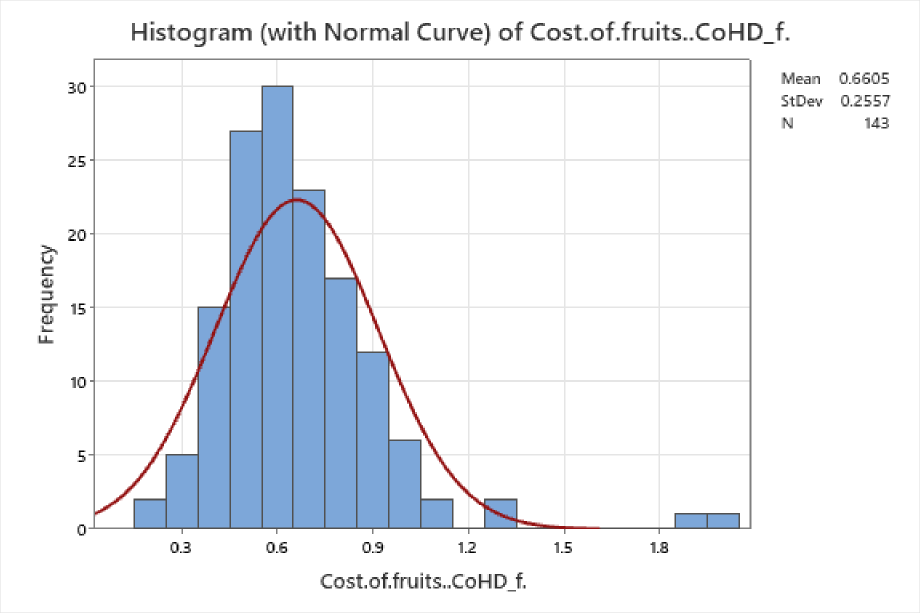


Figure 5: Histogram of Cost of Fruits

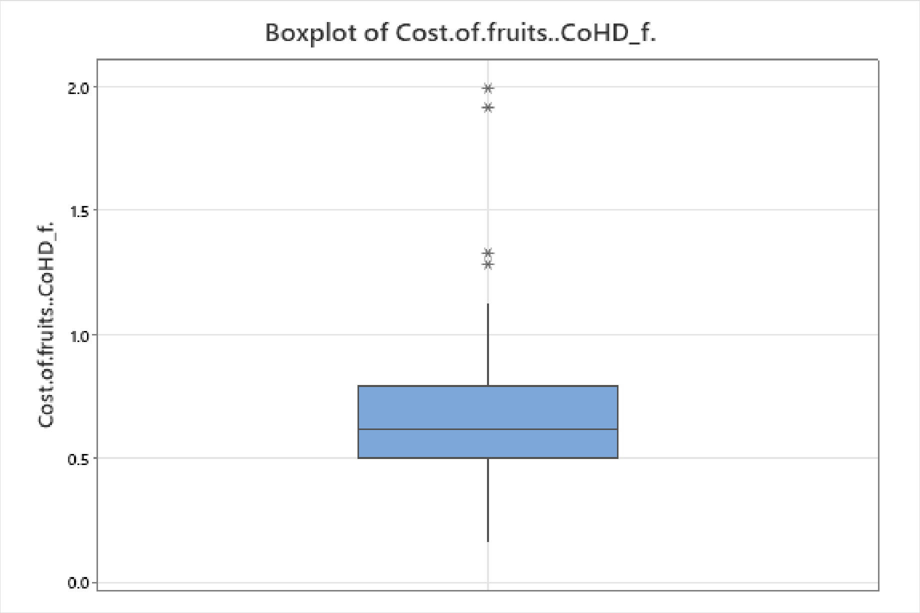


Figure 6: Boxplot of Cost of Fruits

The boxplot clearly indicates a few outliers within the obtained sample. The independent variable, the cost of fruits seems to be also left skewed because of this fact. While most of the countries cost only between $0.3 to $1 for the fruits a person should take per meal, there are a few countries that cost more than $1.8 for that as well.

Cost of Vegetables(independet variable)

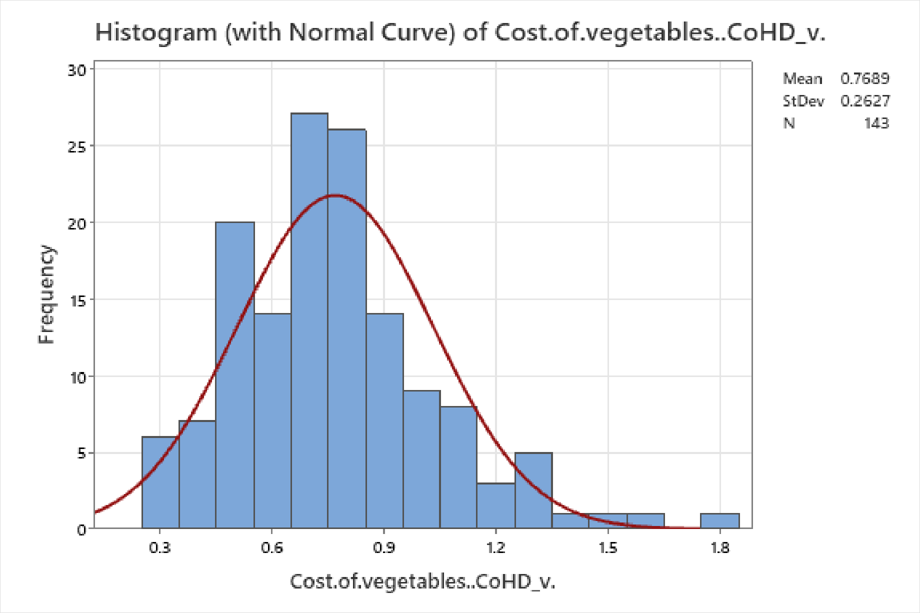


Figure 7: Histogram of Cost of Vegetables

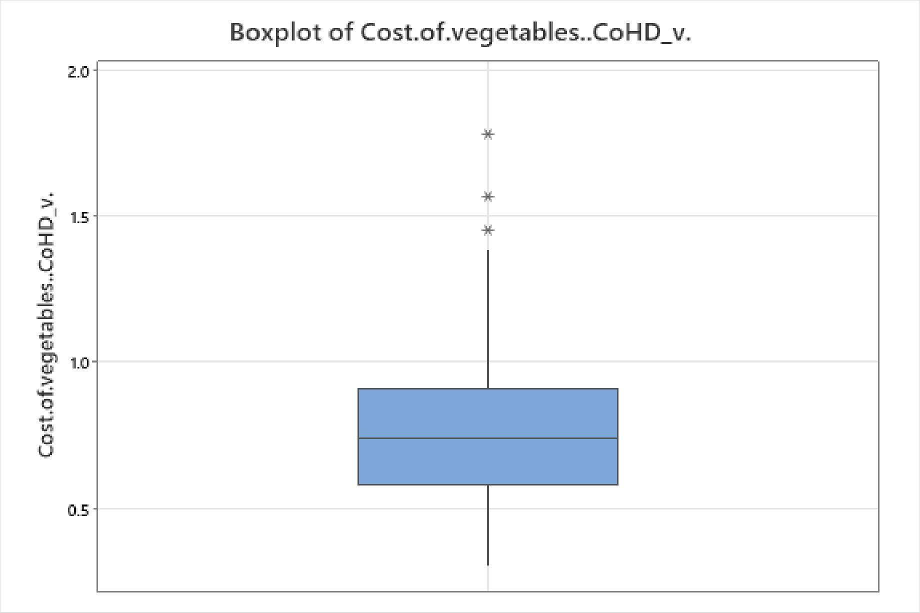


Figure 8: Boxplot of Cost of Vegetables

The data seem to be mostly normally distributed with a few outliers indicated by the boxplot. The figures suggest that the cost of vegetables one should intake in a meal costs around $0.5 to $1 across the counties, with a few deviations.

Cost of Starchy Staples(independet variable)

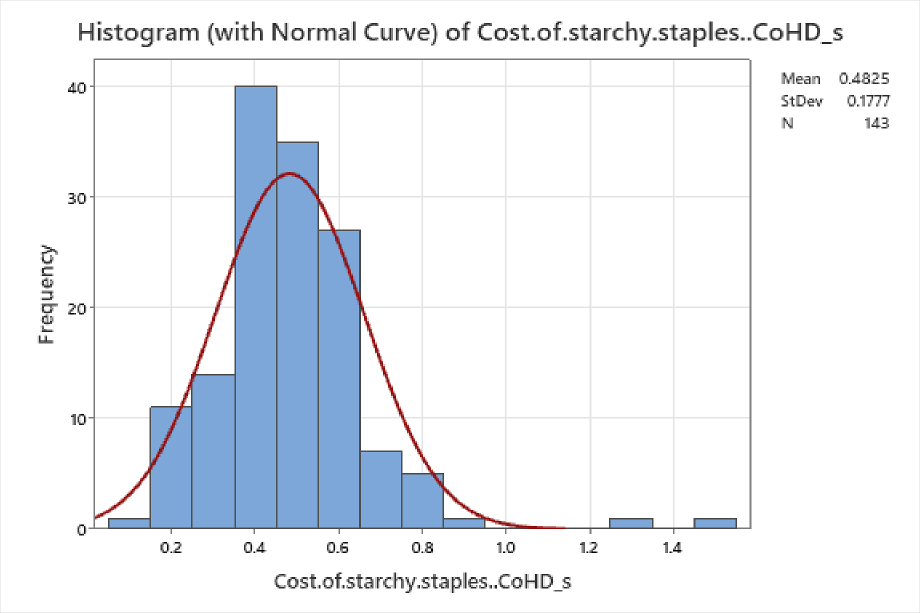


Figure 9: Histogram of Cost of Starchy Staples

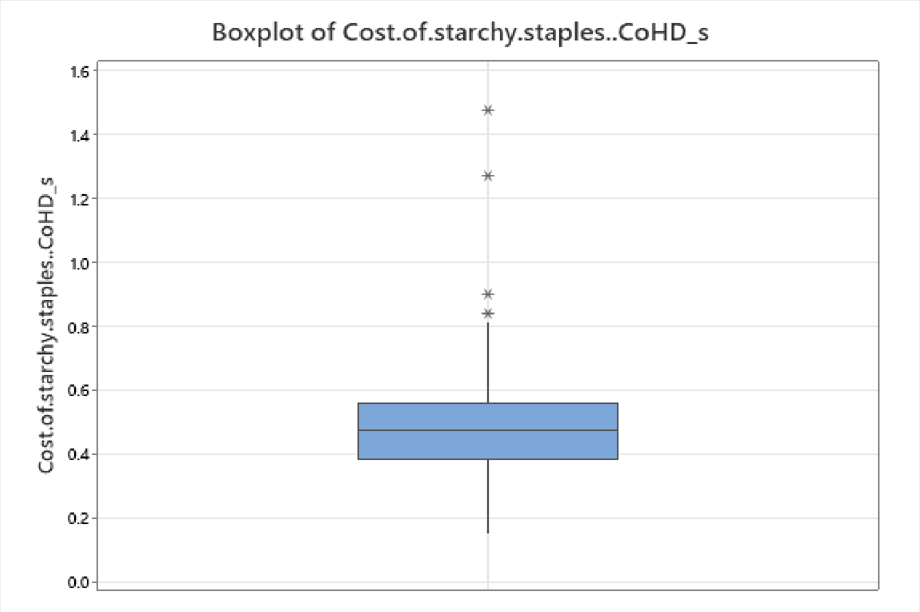


Figure 10: Boxplot of Cost of Starchy Staples

These graphical representations of the cost of starchy staples such as potatoes, rice, and bread which provide carbohydrates vary between \$0.2 to \$1 while some outliers indicated by the boxplot states that there are some countries which cost much higher for them.

Cost of Legumes,Nuts, and Seeds(independet variable)

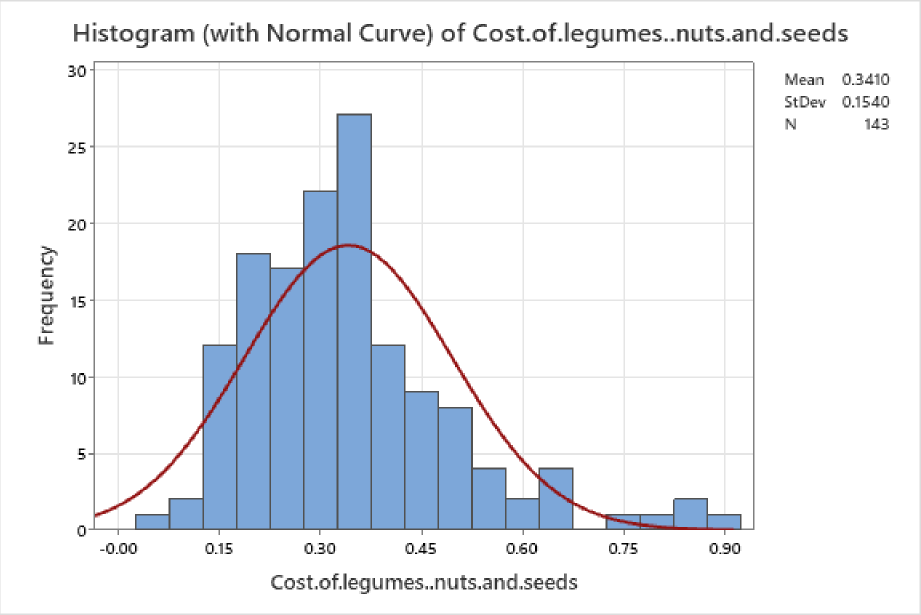


Figure 11: Histogram of Cost of Legumes, Nuts and Seeds

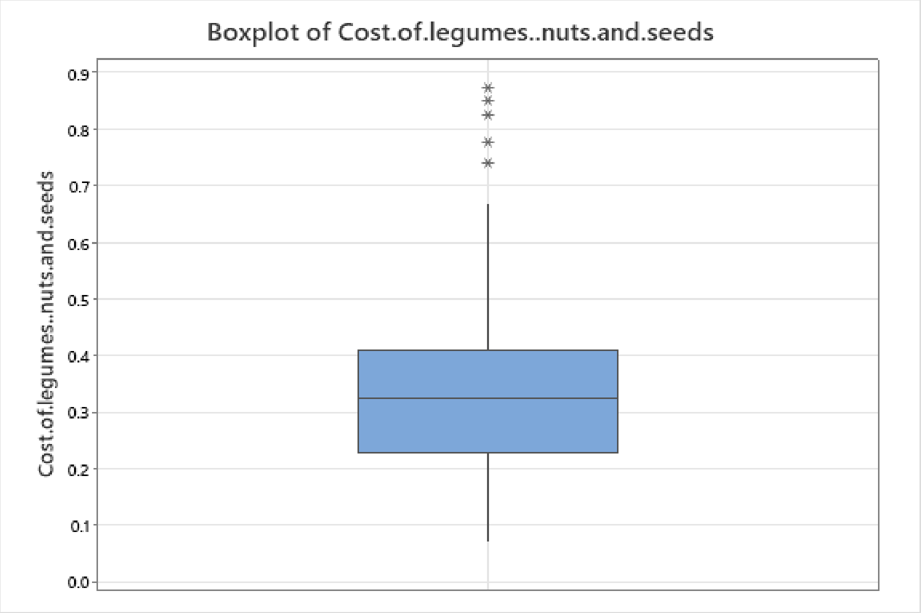


Figure 12: Boxplot of Cost of Legumes, Nuts and Seeds

The boxplot suggests few outliers, while the histogram represents a somewhat skewed but normally distributed data set. Somehow the cost of legumes, nuts, and seeds per meal is kept at less than $1 in all the counties, but less than $0.5 in the majority of them.

Cost of Animal Source Food(independet variable)

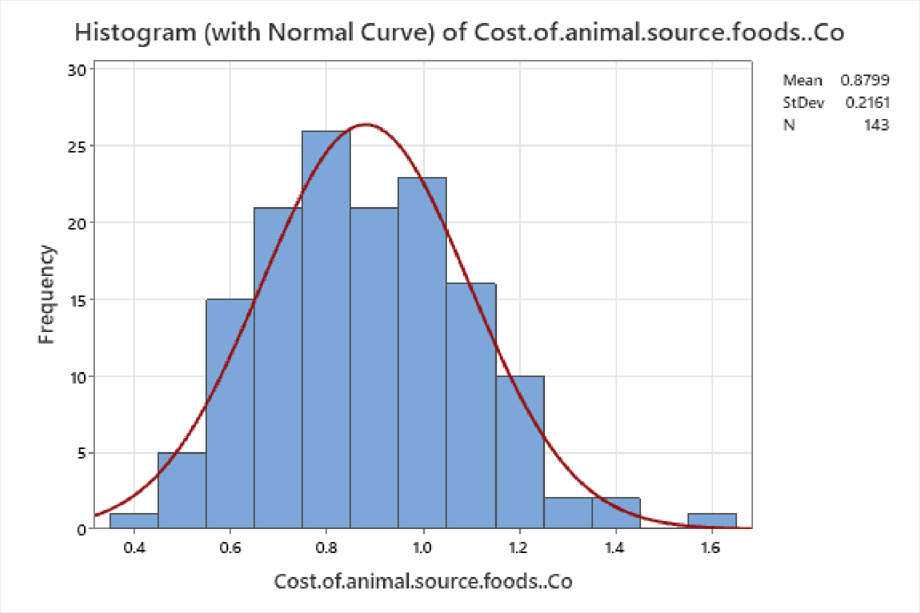


Figure 13: Histogram of Cost of Animal Source Foods

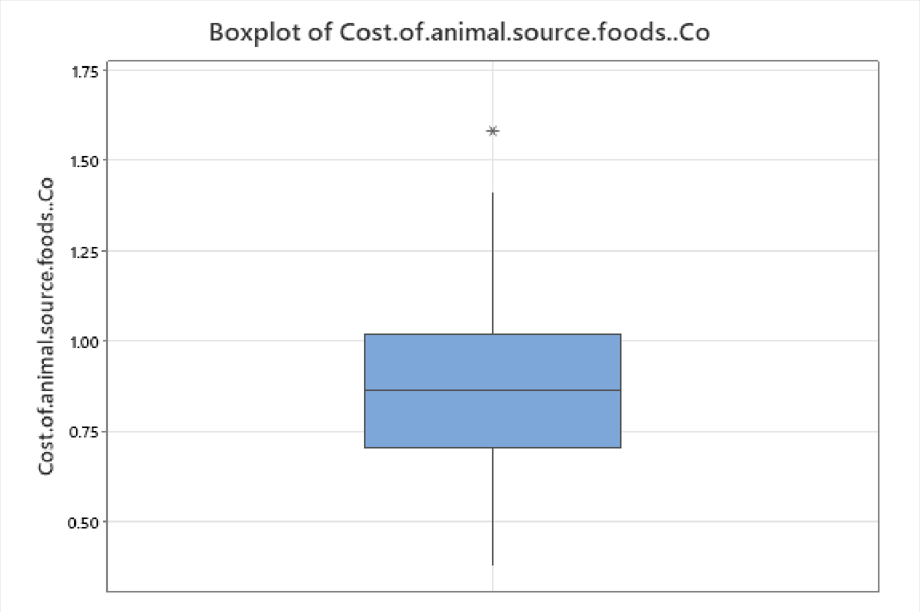


Figure 14: Boxplot of Cost of Animal Source Foods

Both the boxplot and histogram represent a normally distributed dataset with just one outlier. The cost of animal source food around the world varies between $0.5 to $1.4, which is the highest we have seen so far.

Cost of Oils and Fats(independet variable)

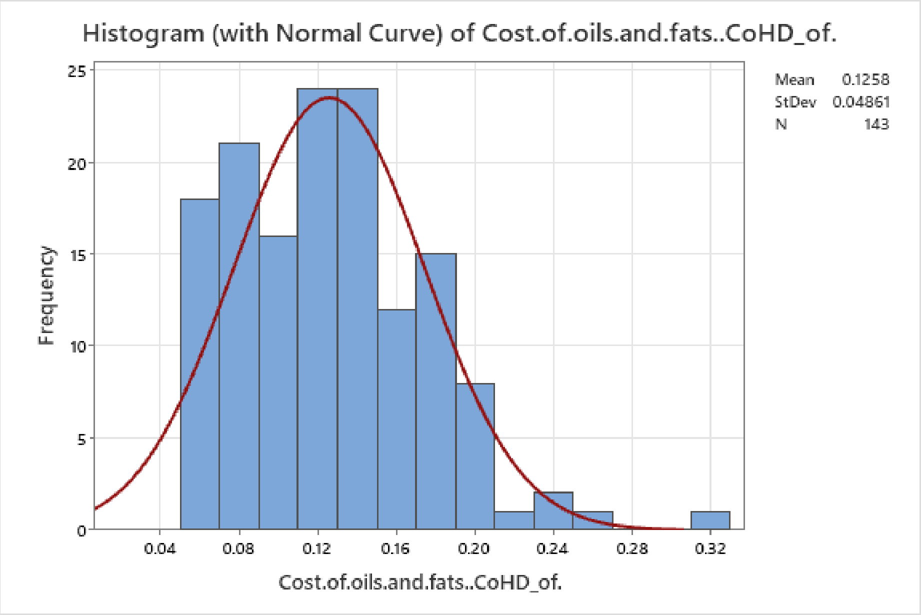


Figure 15: Histogram of Cost of Oils and Fats

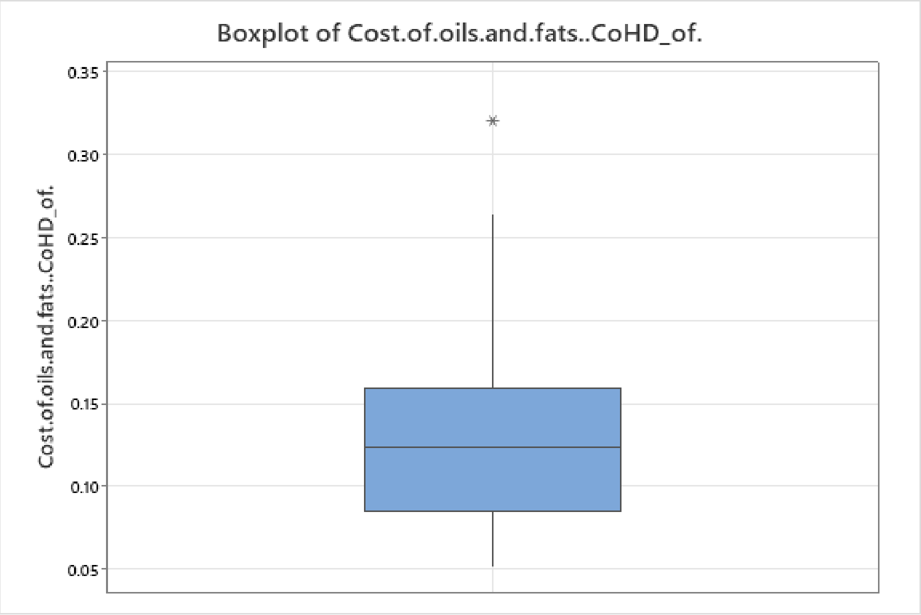


Figure 16: Boxplot of Cost of Oils and Fats

The graphs represent a somewhat normally distributed dataset with only one outlier. The cost of oils and fats remains at less than $0.25 in most countries, which record the lowest among the food types.

Income Range (categorical variable)

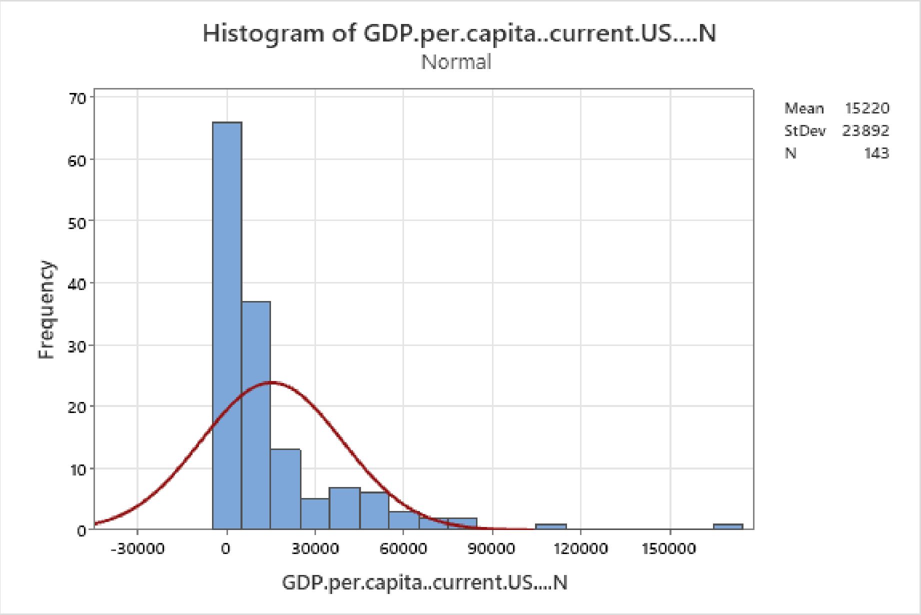


Figure 17: Histogram of GDP per capita

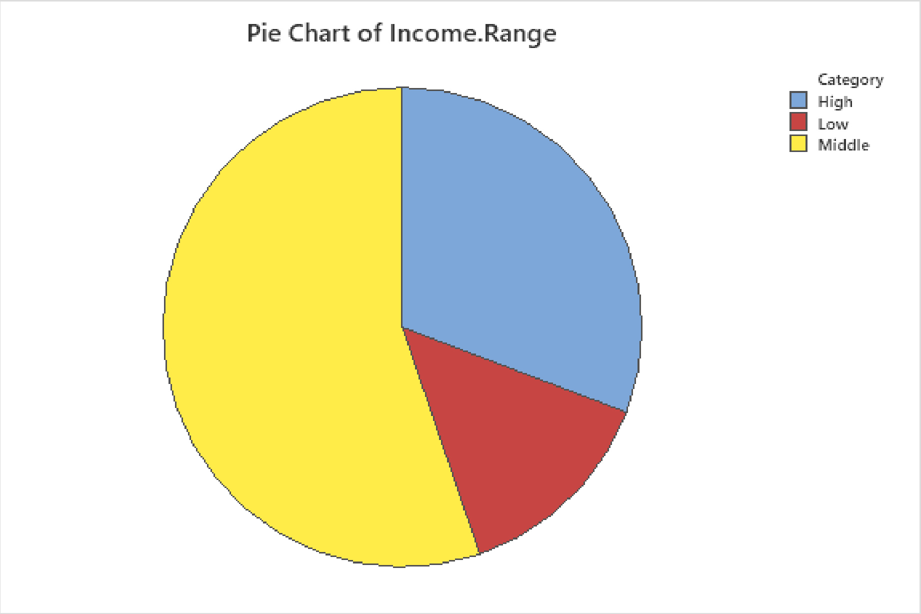


Figure 18: Pie Chart of Income Range

5.2 Quantitative Data Analysis

When starting the data analysis, we face two challenges. First, whether to keep or discard the outliers. Second, fitting the regression model. So we analyzed a few possibilities for the first quest.

1. Attempt on log normalization method for standardizing data.
2. Data analyzation with the outliers.
3. Data analyzation after removing the outliers.

Note that this part only describes the data analysis, the discussion based on that is talked about later in the report.
Then for fitting the regression model, we used stepwise multiple linear regression. From the three possibilities with the outliers, we have decided not to use the data with outliers so when fitting the regression model the dataset with the outliers has been used. We will talk about why we selected that option later in the study.
We have also checked compatibility with the categorical variable (income range) and later discussed that the income range of a country doesn't have a sufficient impact on the cost of an energy-sufficient diet and removed it from the model.

5.2.1 Attempt on log normalization method for standardizing data

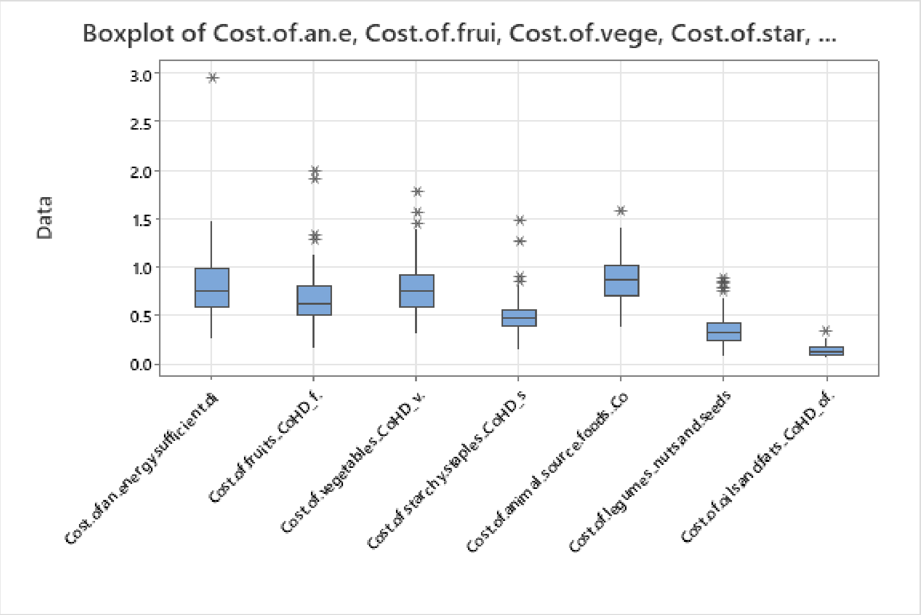


Figure 19: Boxplot of the original sample

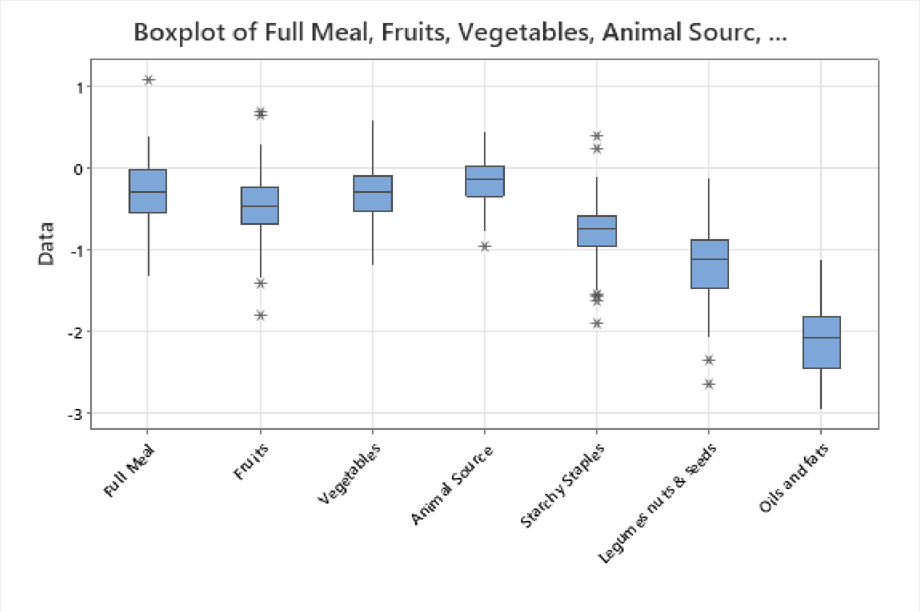


Figure 20: Boxplot after using log normalization

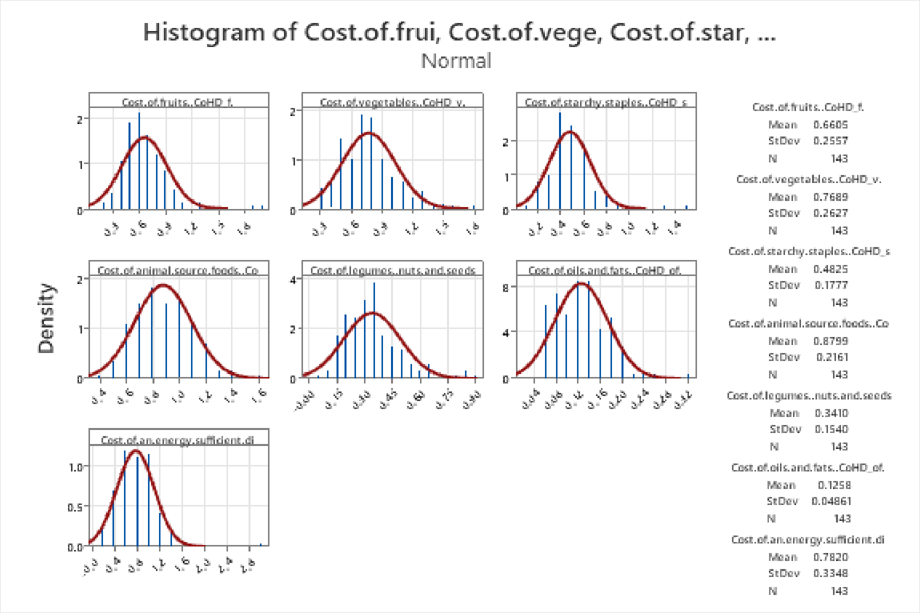


Figure 21: Histograms of the original sample

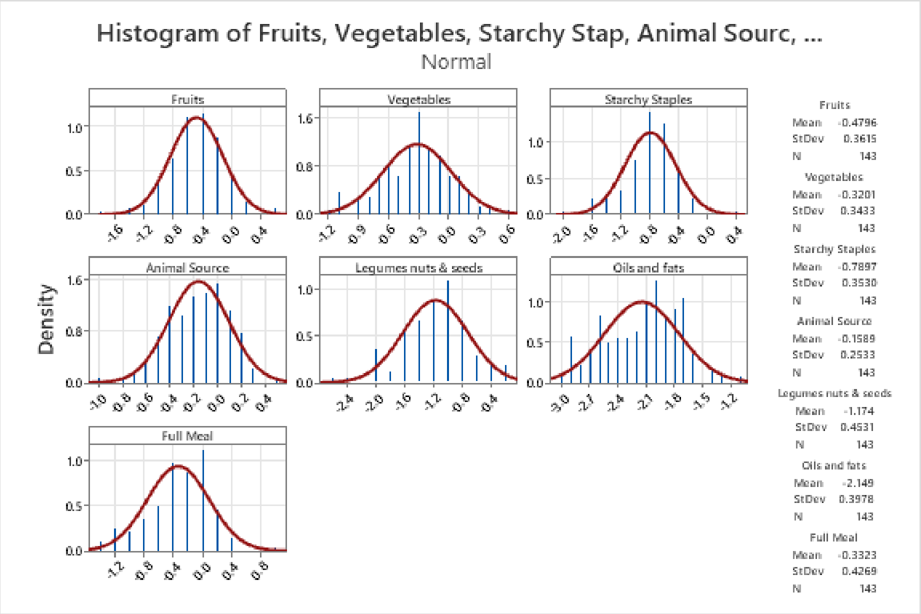
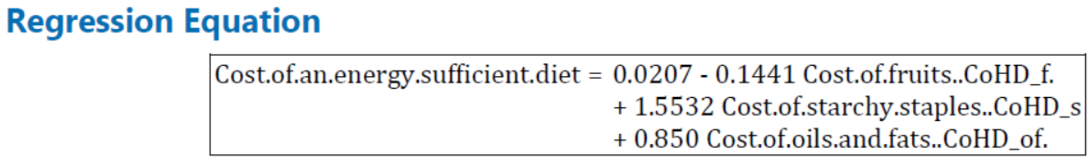


Figure 22: Histograms after using log normalization

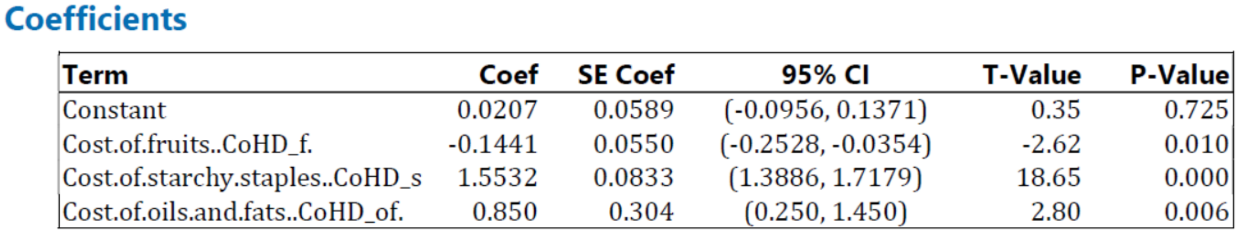
5.2.2 Data Analyzation with the Outliers



Regression Equation

Cost.of.an.energy.sufficient.diet = 0.0207 - 0.1441 Cost.of.fruits..CoHD_f.
+ 1.5532 Cost.of.starchy.staples..CoHD_s
+ 0.850 Cost.of.oils.and.fats..CoHD_of.

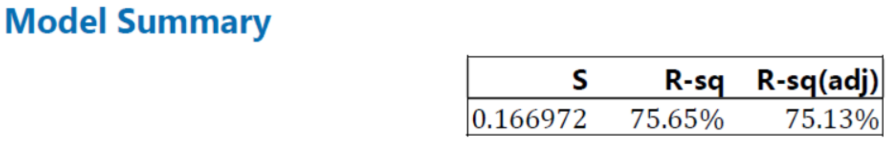
Figure 23: Regression equation with outliers



Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value
Constant	0.0207	0.0589	(-0.0956, 0.1371)	0.35	0.725
Cost.of.fruits..CoHD_f.	-0.1441	0.0550	(-0.2528, -0.0354)	-2.62	0.010
Cost.of.starchy.staples..CoHD_s	1.5532	0.0833	(1.3886, 1.7179)	18.65	0.000
Cost.of.oils.and.fats..CoHD_of.	0.850	0.304	(0.250, 1.450)	2.80	0.006

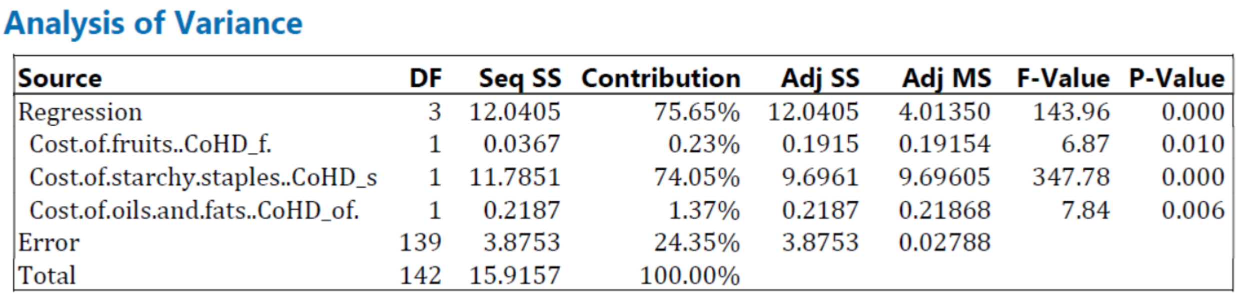
Figure 24: Coefficient table of regression equation with outliers



Model Summary

S	R-sq	R-sq(adj)
0.166972	75.65%	75.13%

Figure 25: Model summary of regression equation with outliers

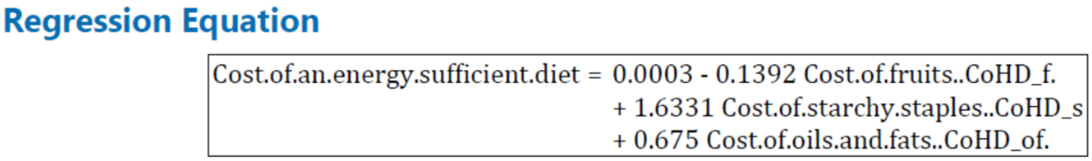


Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	3	12.0405	75.65%	12.0405	4.01350	143.96	0.000
Cost.of.fruits..CoHD_f.	1	0.0367	0.23%	0.1915	0.19154	6.87	0.010
Cost.of.starchy.staples..CoHD_s	1	11.7851	74.05%	9.6961	9.69605	347.78	0.000
Cost.of.oils.and.fats..CoHD_of.	1	0.2187	1.37%	0.2187	0.21868	7.84	0.006
Error	139	3.8753	24.35%	3.8753	0.02788		
Total	142	15.9157	100.00%				

Figure 26: ANOVA table of regression equation with outliers

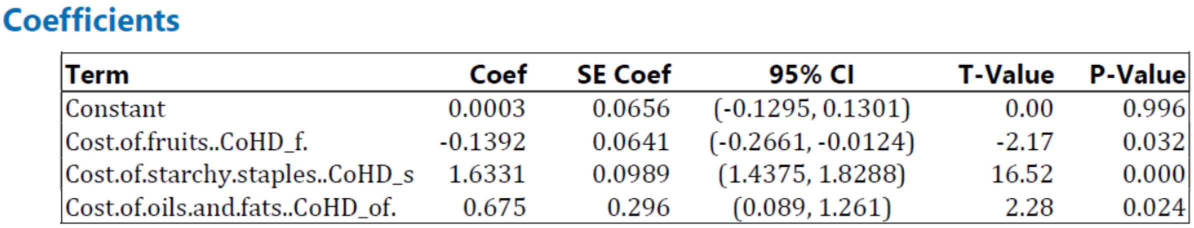
5.2.3 Data Analyzation after removing the Outliers



Regression Equation

Cost.of.an.energy.sufficient.diet = 0.0003 - 0.1392 Cost.of.fruits..CoHD_f.
+ 1.6331 Cost.of.starchy.staples..CoHD_s
+ 0.675 Cost.of.oils.and.fats..CoHD_of.

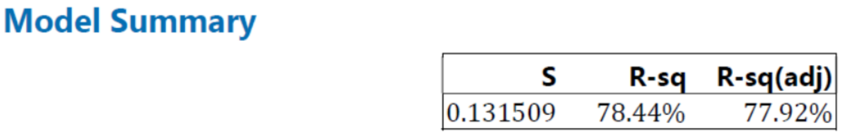
Figure 27: Regression equation after removing outliers



Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value
Constant	0.0003	0.0656	(-0.1295, 0.1301)	0.00	0.996
Cost.of.fruits..CoHD_f.	-0.1392	0.0641	(-0.2661, -0.0124)	-2.17	0.032
Cost.of.starchy.staples..CoHD_s	1.6331	0.0989	(1.4375, 1.8288)	16.52	0.000
Cost.of.oils.and.fats..CoHD_of.	0.675	0.296	(0.089, 1.261)	2.28	0.024

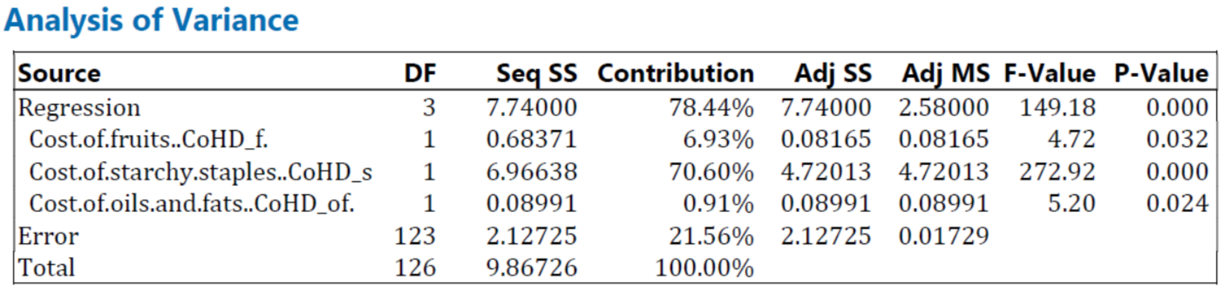
Figure 28: Coefficient table of regression equation after removing outliers



Model Summary

S	R-sq	R-sq(adj)
0.131509	78.44%	77.92%

Figure 29: Model summary of regression equation after removing outliers



Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	3	7.74000	78.44%	7.74000	2.58000	149.18	0.000
Cost.of.fruits..CoHD_f.	1	0.68371	6.93%	0.08165	0.08165	4.72	0.032
Cost.of.starchy.staples..CoHD_s	1	6.96638	70.60%	4.72013	4.72013	272.92	0.000
Cost.of.oils.and.fats..CoHD_of.	1	0.08991	0.91%	0.08991	0.08991	5.20	0.024
Error	123	2.12725	21.56%	2.12725	0.01729		
Total	126	9.86726	100.00%				

Figure 30: ANOVA table of regression equation after removing outliers

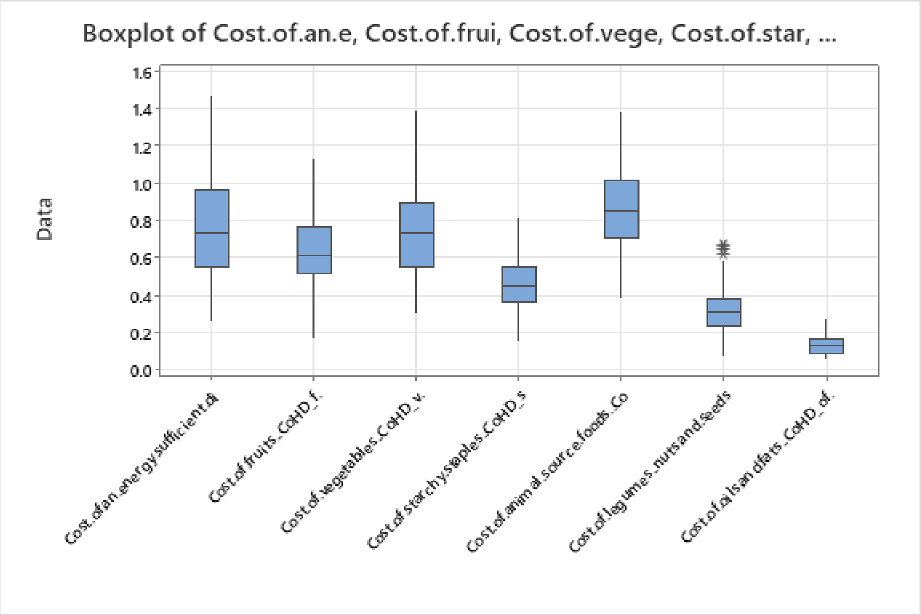


Figure 31: Boxplot after removing outliers once

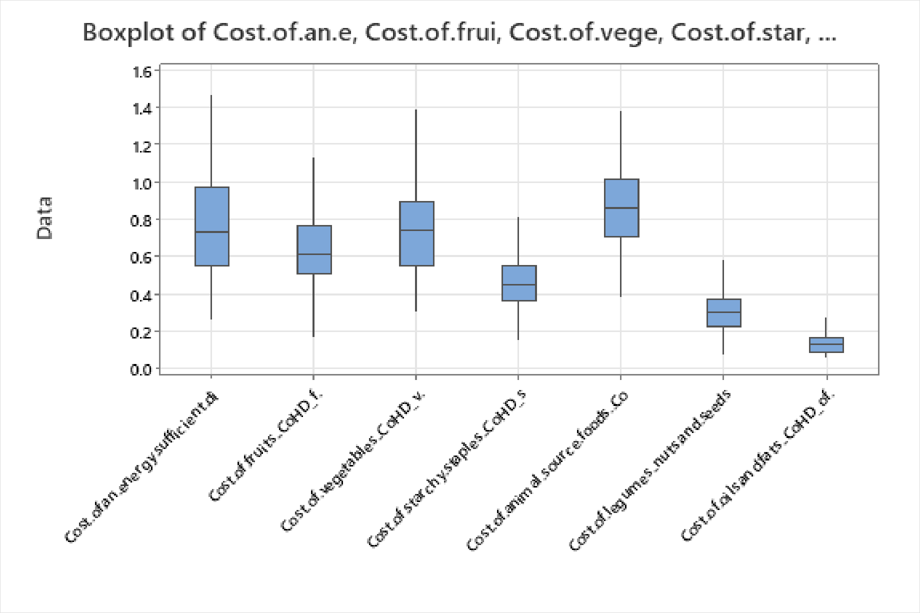


Figure 32: Boxplot after removing outliers twice

5.2.4 Study of categorical data

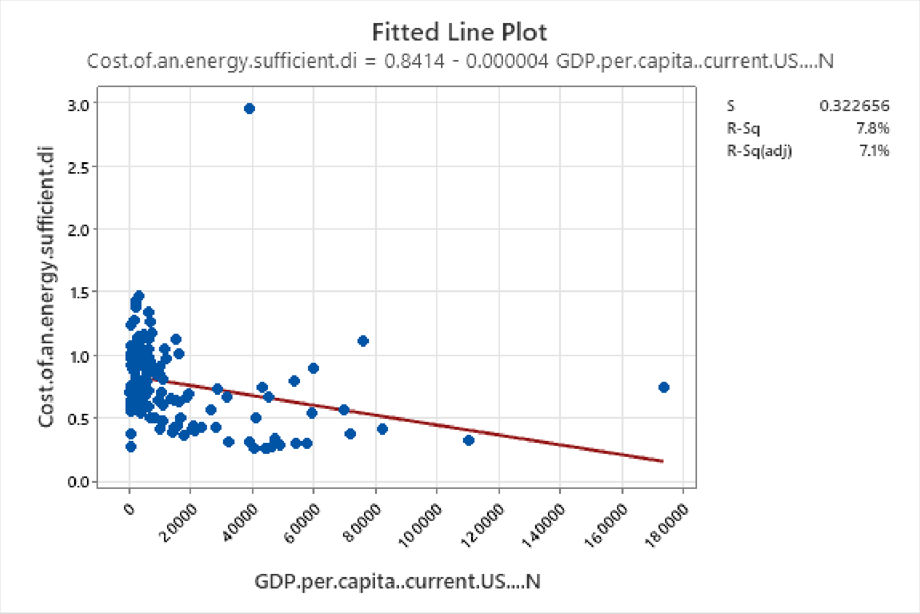


Figure 33: Scatterplot of Cost of an energy-sufficient meal vs GDP per capita

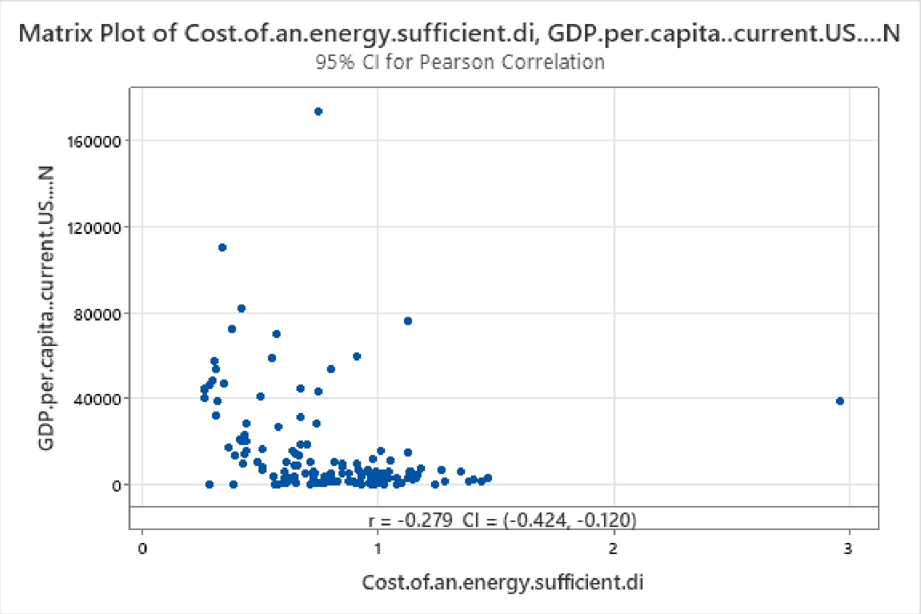
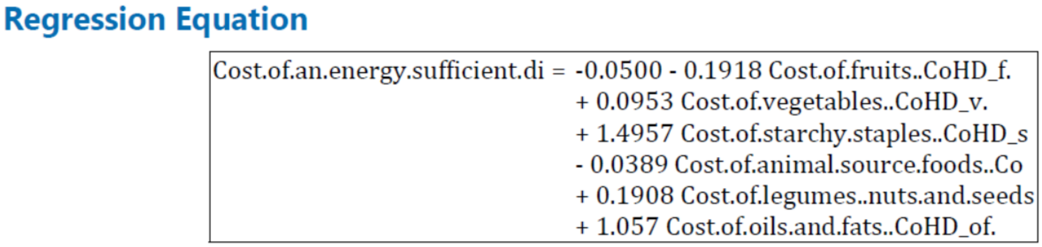


Figure 34: Correlation between Cost of an energy-sufficient meal and GDP per capita

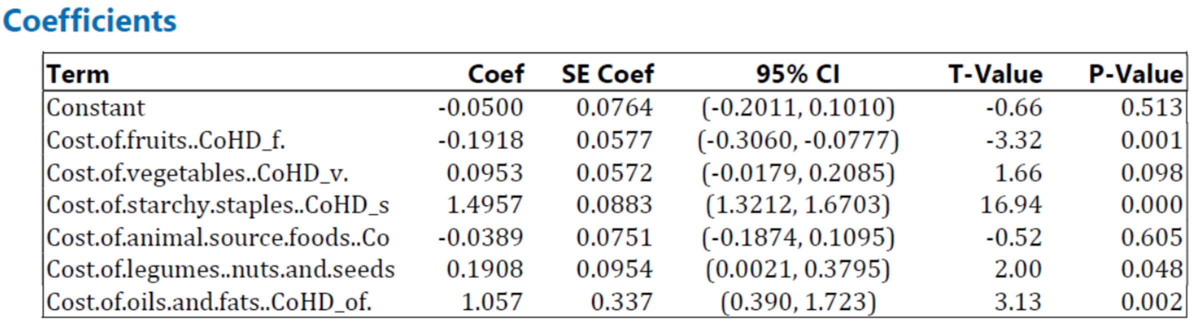
5.2.5 Initial model



Regression Equation

```
Cost.of.an.energy.sufficient.di = -0.0500 - 0.1918 Cost.of.fruits..CoHD_f.
                                  + 0.0953 Cost.of.vegetables..CoHD_v.
                                  + 1.4957 Cost.of.starchy.staples..CoHD_s
                                  - 0.0389 Cost.of.animal.source.foods..Co
                                  + 0.1908 Cost.of.legumes..nuts.and.seeds
                                  + 1.057 Cost.of.oils.and.fats..CoHD_of.
```

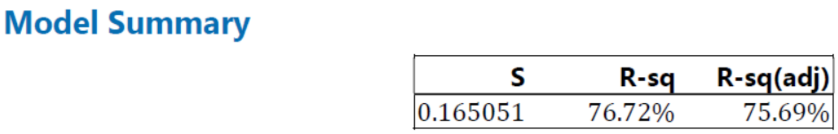
Figure 35: Regression equation of the initial model



Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value
Constant	-0.0500	0.0764	(-0.2011, 0.1010)	-0.66	0.513
Cost.of.fruits..CoHD_f.	-0.1918	0.0577	(-0.3060, -0.0777)	-3.32	0.001
Cost.of.vegetables..CoHD_v.	0.0953	0.0572	(-0.0179, 0.2085)	1.66	0.098
Cost.of.starchy.staples..CoHD_s	1.4957	0.0883	(1.3212, 1.6703)	16.94	0.000
Cost.of.animal.source.foods..Co	-0.0389	0.0751	(-0.1874, 0.1095)	-0.52	0.605
Cost.of.legumes..nuts.and.seeds	0.1908	0.0954	(0.0021, 0.3795)	2.00	0.048
Cost.of.oils.and.fats..CoHD_of.	1.057	0.337	(0.390, 1.723)	3.13	0.002

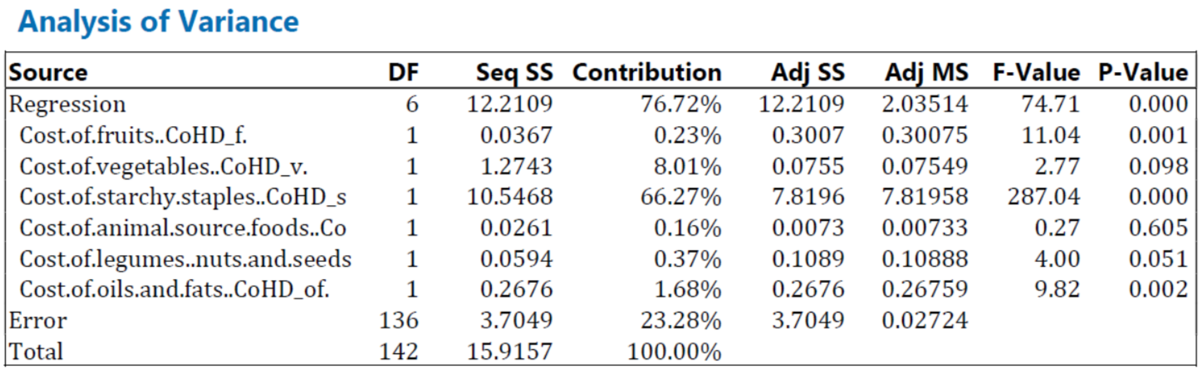
Figure 36: Coefficient table of regression equation of the initial model



Model Summary

S	R-sq	R-sq(adj)
0.165051	76.72%	75.69%

Figure 37: Model summary of regression equation of the initial model



Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	6	12.2109	76.72%	12.2109	2.03514	74.71	0.000
Cost.of.fruits..CoHD_f.	1	0.0367	0.23%	0.3007	0.30075	11.04	0.001
Cost.of.vegetables..CoHD_v.	1	1.2743	8.01%	0.0755	0.07549	2.77	0.098
Cost.of.starchy.staples..CoHD_s	1	10.5468	66.27%	7.8196	7.81958	287.04	0.000
Cost.of.animal.source.foods..Co	1	0.0261	0.16%	0.0073	0.00733	0.27	0.605
Cost.of.legumes..nuts.and.seeds	1	0.0594	0.37%	0.1089	0.10888	4.00	0.051
Cost.of.oils.and.fats..CoHD_of.	1	0.2676	1.68%	0.2676	0.26759	9.82	0.002
Error	136	3.7049	23.28%	3.7049	0.02724		
Total	142	15.9157	100.00%				

Figure 38: ANOVA table of regression equation of the initial model

6 Discussion and Conclusion

6.1 Discussion

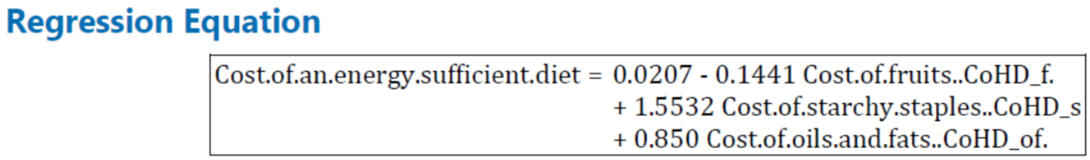
After getting rid of the missing values problem, the next challenge we faced was the outliers. As you can see in the boxplot in “Figure 19”, we have found that there were some outliers amongst the data. Faced with this challenge we were given with few options, i.e.: Standardizing data, removing the outliers, or accepting them as relevant data in spite of being graphically represented as outliers.

First, we tried standardizing the data using the log normalization method. We converted the original data into their natural logarithm values, expecting them to be normalized. But as you can see in “Figure 20”, that method introduced new outliers and made the situation worse. Furthermore, according to “Figures 21 and 22” the histograms show that the distributions haven’t been affected much by this action. Thus, we decided that this method was not suitable for this case.

This outcome made us wonder if the outliers are truly outliers or not. So, we decided to fit them into regression models and check their compatibility with the models. We used stepwise method with 0.05 significance level to fit the models.

In both cases (analyzing with and without outliers) our predictor variable count dropped from six to three. The remaining independent variables which affect the cost of an energy-sufficient diet were the cost of fruits, the cost of starchy staples, and the cost of oils and fats. According to “Figure 23” and “Table 01”, the coefficients of these variables are -0.1441, 1.5532, and 0.850 respectively when we analyze the data with the outliers. And the constant is 0.0207. Next, when we analyze the data after removing the outliers, the coefficients are -0.1392, 1.6331, and 0.675 respectively. And the constant is 0.0003. Meaning, the regression equations are

With outliers:

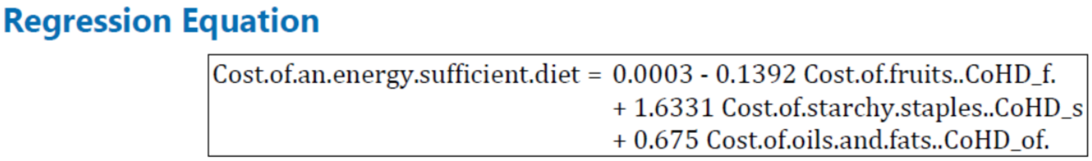


Regression Equation

```
Cost.of.an.energy.sufficient.diet = 0.0207 - 0.1441 Cost.of.fruits..CoHD_f.
                                    + 1.5532 Cost.of.starchy.staples..CoHD_s
                                    + 0.850 Cost.of.oils.and.fats..CoHD_of.
```

Figure 39: Regression Equation with Outliers

Without outliers:



Regression Equation

```
Cost.of.an.energy.sufficient.diet = 0.0003 - 0.1392 Cost.of.fruits..CoHD_f.
                                    + 1.6331 Cost.of.starchy.staples..CoHD_s
                                    + 0.675 Cost.of.oils.and.fats..CoHD_of.
```

Figure 40: Regression Equation after removing Outliers

And the R-squared values of the models with outliers and after removing outliers are 75.65% and 78.44%. Which is an increase of 2.79%. Also, the coefficients are not significantly changed. The differences aren't enough to choose one method over the other. If one were to choose one - 29 - method, we recommend choosing the method where we don't remove the outliers. The reason would be, that even if the data shows graphically as outliers, they might not actually be outliers. Bad sampling methods and smaller samples may cause this. And even more, removing 20 data because of an outlier in one or two predictor variables of that data row would be pointless for such a small change in the accuracy of the model. So, without damaging the dataset, taking tho original data with the outliers would be the best option in this case.

6.1.1 Hypothesis Analysis

Hypothesis 1 - Regression model without removing outliers

Null Hypothesis (H_0) = The model is adequate.
Alternative Hypothesis (H_a) = The model is inadequate.
Test Statistic:

- ANOVA table of regression equation with outliers (26)
- ANOVA table of regression equation of the initial model (38)

The following formula is used to prove the test statistic

$$F_{Partial} = \frac{[SS_{Reg}^{Full} - SS_{Reg}^{Reduced}]/(k-p)}{SS_{Error}^{Full}/(n-(k+1))}$$

$$= \frac{(12.2109-12.0405)/(6-3)}{3.7049/(143-(6+1))}$$

$$= 2.0850$$

$$F_{Table,0.05,137,136} = 1.32663$$

Figure 41: Partial F-value calculation for hypothesis 01

Thus,
$F_{\text{Partial}} > F_{\text{Table}}$
Conclusion: Reject null hypothesis. The model is inadequate.

Hypothesis 2 - Regression model after removing outliers

Null Hypothesis (H$_0$) = The model is adequate.
Alternative Hypothesis (H$_a$) = The model is inadequate.
Test Statistic:

- ANOVA table of regression equation after removing outliers (30)
- ANOVA table of regression equation of the initial model (38)

The following formula is used to prove the test statistic

$$F_{Partial} = \frac{[SS_{Reg}^{Full} - SS_{Reg}^{Reduced}]/(k-p)}{SS_{Error}^{Full}/(n-(k+1))}$$

$$= \frac{(12.2109-7.7400)/(6-3)}{3.7049/(123-(6+1))}$$

$$= 46.6611$$

$$\mathrm{F}_{\mathrm{Table},0.05,117,116} = 1.3581$$

Figure 42: Partial F-value calculation for hypothesis 02

Thus,
$\mathrm{F}_{\mathrm{Partial}} > \mathrm{F}_{\mathrm{Table}}$
Conclusion: Reject null hypothesis. The model is inadequate.

Hypothesis 3 – Coefficient values of of the reduced model with outliers

Null Hypothesis (H$_0$) = The cost of an energy-sufficient diet does not significantly depend on the cost of fruits, cost of vegetables, cost of starchy staples, cost of animal-source foods, cost of legumes, nuts, and seeds, and cost of oils, and fats.
Alternative Hypothesis (H$_a$) = The cost of an energy-sufficient diet significantly depends on at least one of the independent variables. (Cost of fruits, cost of vegetables, cost of starchy staples, cost of animal-source foods, cost of legumes, nuts, and seeds, and cost of oils, and fats)

$$H_0 : \beta_i = 0, H_1 : \beta_i \neq 0 \tag{1}$$

Test Statistic: ANOVA table of regression equation of the initial model (38)

$$\beta_1 = (CostofFruits) \tag{2}$$

P value : $0.001 < 0.05$
Conclusion : Reject Null Hypothesis
$\beta_1 \neq 0$
This confirms the coefficient value of the cost of fruits given by the stepwise method, $\beta_1 = -0.44$

$$\beta_2 = (CostofVegetables) \tag{3}$$

P value : $0.098 > 0.05$
Conclusion : Do not reject Null Hypothesis
$\beta_2 = 0$
This confirms the coefficient value of the cost of vegetables given by the stepwise method,
$\beta_2 = 0$

$$\beta_3 = (CostofStarchyStaples) \tag{4}$$

P value : $0.000 < 0.05$
Conclusion : Reject Null Hypothesis
$\beta_3 \neq 0$
This confirms the coefficient value of the cost of starchy staples given by the stepwise method,
$\beta_3 = 1.5532$

$$\beta_4 = (Costofanimal - sourcefood) \tag{5}$$

P value : $0.605 > 0.05$
Conclusion : Do not reject Null Hypothesis
$\beta_4 = 0$
This confirms the coefficient value of the cost of animal source food given by the stepwise method,
$\beta_4 = 0$

$$\beta_5 = (Costoflegumes, nuts, andseeds) \tag{6}$$

P value : $0.098 > 0.05$
Conclusion : Do not reject Null Hypothesis
$\beta_5 = 0$
This confirms the coefficient value of the cost of legumes, nuts, and seeds given by the stepwise method, β
0

$$\beta_6 = (Costofoilsandfats) \tag{7}$$

P value : $0.002 < 0.05$
Conclusion : Reject Null Hypothesis
$\beta_6 \neq 0$
This confirms the coefficient value of the cost of oils and fats given by the stepwise method, $\beta_6 = 0.850$

Hypothesis 4 - Coefficient values of the initial model

Null Hypothesis (H_0) = The cost of an energy-sufficient diet does not significantly depend on the cost of fruits, cost of vegetables, cost of starchy staples, cost of animal-source foods, cost of legumes, nuts, and seeds, and cost of oils, and fats.
Alternative Hypothesis (H_a) = The cost of an energy-sufficient diet significantly depends on at least one of the independent variables. (Cost of fruits, cost of vegetables, cost of starchy staples, cost of animal-source foods, cost of legumes, nuts, and seeds, and cost of oils, and fats)

$$H_0 : \beta_i = 0, H_1 : \beta_i \neq 0 \tag{8}$$

Test Statistic: ANOVA table of regression equation of the initial model (38)

$$\beta_1 = (CostofFruits) \tag{9}$$

P value : $0.001 < 0.05$
Conclusion : Reject Null Hypothesis
$\beta_1 \neq 0$
This confirms the coefficient value of the cost of fruits given by the stepwise method, $\beta_1 = -0.1918$

$$\beta_2 = (CostofVegetables) \tag{10}$$

P value : $0.098 > 0.05$
Conclusion : Do not reject Null Hypothesis
$\beta_2 = 0$
This states that the coefficient value of the cost of vegetables given by the stepwise method, $\beta_2 = 0$, contradictory to the model.

$$\beta_3 = (CostofStarchyStaples) \tag{11}$$

P value : $0.000 < 0.05$
Conclusion : Reject Null Hypothesis
$\beta_3 \neq 0$
This confirms the coefficient value of the cost of starchy staples given by the stepwise method,

$\beta_3 = 1.4957$

$$\beta_4 = (Costofanimal - sourcefood) \tag{12}$$

P value : $0.605 > 0.05$
Conclusion : Do not reject Null Hypothesis
$\beta_4 = 0$
This confirms the coefficient value of the cost of animal-source food given by the stepwise method,
$\beta_4 = 0$

$$\beta_5 = (Costoflegumes, nuts, andseeds) \tag{13}$$

P value : $0.098 > 0.05$
Conclusion : Do not reject Null Hypothesis
$\beta_5 = 0$
This confirms the coefficient value of the cost of legumes, nuts, and seeds given by the stepwise method,
$\beta_5 = 0$

$$\beta_6 = (Costofoilsandfats) \tag{14}$$

P value : $0.002 < 0.05$
Conclusion : Reject Null Hypothesis
$\beta_6 \neq 0$
This confirms the coefficient value of the cost of oils and fats given by the stepwise method,
$\beta_6 = 0.1057$

6.2 Conclusion

By all means, in every model the cost of fruits had a negative impact on the cost of an energysufficient meal, while the cost pf oils and fats, and the cost of starchy staples has an positive impact. Cost of starchy staples seems to have a very large coefficient on the regression model in every case.
In conclusion, we can say that reducing the price of starchy staples will affect the cost of an energy-sufficient meal to go down. Though it seems that increasing the price of fruits will reduce the cost of an energy-sufficient diet, it is a question if it's practically true or not. Maybe the calorie intake is not related to fruits and very much related to starchy staples.
Overall, the study provides understanding of various connections between the pricing of different food types and the cost of an energy-sufficient diet. These findings will be helpful when managing health and nutrition values of a meal across the globe as well as making policies and decision related to food pricing targeting the wellbeing of the people. Such activities will be critical in ensuring global food security, better dietary patterns, and improved health outcomes.

7 Summary

The study follows the relationship between the cost of various food types and the cost of an energy-sufficient diet. We have collected relevant data from a well-established foundation, the World Data Bank and followed the standards of food nutrition recommended by the World Health Organization. Using exploratory data analysis to visualize the collected data and following well famed theories to cleanup and process the data using multiple linear regression, the study continued. Using hypotheses to test predicted outcomes, the study came to a conclusion that starchy staples has the most effect in deciding the price of an energy-sufficient diet.

8 References

- Book: Food and Agriculture Organization of the United Nations: Cost and affordability of healthy diets across and within countries, Background paper for the State of Food Security and Nutrition in the World 2020.

- Book: Food and Agriculture Organization of the United Nations: The State of Food Security and Nutrition in the World, Transforming Food Systems for Affordable Healthy Diets.

- https://www.who.int/

- https://scholar.google.com/

- https://databank.worldbank.org/source/food prices-for-nutrition

- https://databank.worldbank.org/source/world-development-indicators

- https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-incomelevel-2021-2022

9 Appendix

R code for data cleaning process

```
# Load the necessary library for data manipulation
library(dplyr)
# Read the CSV file into a data frame
df <- read.csv(”data.csv”)
# Check for missing values in each variable
missing_values <- sapply(df, function(x) sum(is.na(x)))
# Print the number of missing values for each variable
print(missing_values)
# Remove rows with missing values
df_clean <- na.omit(df)
# Store the cleaned data in a new CSV file
write.csv(df_clean, file = ”cleaned_data.csv”, row.names = FALSE)
```