Dollar General

Email: tvasos@dollargeneral.com
Website: https://www.dollargeneral.com/

BUSINESS REPORT

Kevin Sanagustin, Monica Luevano, Adam Luong,
Data Analysts
Ksanagu@calstatela.edu mluevan6@calstatela.edu aluong35@calstatela.edu

TABLE OF CONTENTS

Executive Summary

Dollar General is one of the most sought-after extreme value retailers in the United States market. Under CEO David Perdue's supervision, the company saw exponential growth, with revenues reaching $9.2 billion. Our project aims to maintain Dollar General's growth by guiding the company to expand to more locations where its targeted audience can be found using health data. This project will specifically dive into international data as Dollar General's presence is not as strong as it can be internationally. The project will provide countries where Dollar General can expand to and keep its growth in profits.

Since its founding, Dollar General has prominently pursued its mission of providing affordable everyday essentials to low—and middle-income consumers. Dollar General operates in 35 states, primarily in rural and suburban areas. Their operations in rural and suburban areas have given them the upper hand against large retailers who choose to operate elsewhere. However, while the success of Dollar General is undeniable, they face strategic challenges of how to optimize their opportunities to grow.

Challenges the company faced were merchandising and operational improvements, service expansion, store formats, and international expansion. These challenges steered Dollar General to look for what was the company's next venture. Addressing any of these challenges would have a positive impact on the company's growth. Navigating through the number of paths Dollar General can take is essential when deciding which of these problems is most important to assess to be on a sustainable path to continue to be a top contender of greatest extreme value retailers.

With a solid market presence and strong financial presence, Dollar General faces the decision on how to maintain their consistent growth. Choosing which challenges to address and focus on will determine the success of the company. Dollar General's CEO's decision will be crucial in navigating these opportunities to sustain growth and keep Dollar General's competitive edge.

Over the past decades, Dollar General has conquered rural America by bringing affordable items to low-income communities. While the company has been successful over time, it is still important to do a self-assessment. A SWOT analysis helps determine what areas a company can improve or utilize in order to tailor our strategic goals. The SWOT analysis goes over a company's strengths, weaknesses, opportunities, and threats. This project will utilize this analysis to help Dollar General improve in one of the areas of the SWOT analysis.

The strengths at Dollar General are that 80% of these stores are in low populated rural areas across America. This allows them to get customers that are far away from big retailers such as Walmart. The population in these parts sell mostly convenience, grocery, and drug merchant items that benefit and are sought out by these types of communities.

The weakness of Dollar General is that all stores are only limited to the U.S and have no presence outside, like its competitor PriceSmart, however, it is unsure if expanding would result in regional instability and inability to uphold business. Additionally, there are concerns based on nutritional factors Dollar General currently provides.

Furthermore, Dollar General is also limited in stores in the United States. Dollar General is mostly found in the Southeast United States. The company isn't widening its scope of customers by opening more stores outside of the U.S.

The opportunity is an increase in demand for private label products. There has been a recent preference towards private label products because they are cheaper than established name brands and while improving the quality.

The threats are that Dollar General is dependent on affordable prices but with demands in labor due to demands in higher wages and economic recession period. Additionally, the company regulates under U.S law that are constantly changing that add significant expenses.

After reviewing the SWOT analysis, this project will focus on Dollar General's weakness of having limited stores internationally and low-quality nutritional options. Specifically, this project will investigate the expansion of Dollar General stores outside of the United States and into other regions based on nutritional data. The dataset on foreign regions, like Africa and Asia, will be used to determine where Dollar General can best implement stores to further their consistent business growth.

Business Understanding

Business Objectives

- Provide everyday essentials to international locations.
- Locate where their targeted audience is located to expand their stores.

Business Success Criteria

- Improve Dollar General's global expansion by identifying the most suitable countries based on a combination of BMI, nutritional deficiencies, and environmental factors.

Assess the Situation

Risks and contingencies

- Risks that may occur are missing values, excessive fields, inaccurate values, outliers, and not enough regions.
- If risks occur, we will have inaccurate calculations, unnecessary fields/numbers, incorrect analysis, biased calculated outcomes, and low information on specific regions that lead to poor quality analysis and insights in our business goals.
- If risks listed above occur, we need to have a cleaning data process to remove errors, make tables for necessary fields only, making box plots to locate outliers, doing correction methods on values, and brainstorm adjusting our analysis based on data.

Data Analysis Goals

- Data analysis goals by problem type:
    - Description
        - Identify countries with the highest BMI (a proxy for obesity) within Asia, North America, and Latin America
        - Analyze daily fruit and vegetable intake in those countries to assess nutritional deficiencies
        - Analyze average planetary boundary impact from freshwater use in regions
    - Dependency
        - Determine if there is a correlation between BMI and gender
    - Prediction
        - Assess the relationship between environmental impacts and obesity

Data Understanding

Collect Initial Data

- The following dataset is the global nutrition report from 2021 for all countries.

Table 1: **This table shows our inclusion and exclusion criteria for the global nutrition report dataset.**

Variable Name	Included/Excluded	Rational
iso3	Included	Relevant three-letter code for country, used to connect tables
country	Included	Relevant to identify location
disaggregation	Included	Relevant for state what we are identifying
disagg.value	Included	Relevant for identifying sex
region	Included	Relevant for identifying region
subregion	Included	Relevant for identifying subregion
section	Included	Relevant for identifying section
obesity	Included	Relevant for identifying obesity based on BMI
year	Included	Relevant for identifying year
Fruit	Included	Relevant for identifying low nutrition intake
Vegetables	Included	Relevant for identifying low nutrition intake
Legumes	Included	Relevant for identifying low nutrition intake
Nuts	Included	Relevant for identifying low nutrition intake
Whole grains	Included	Relevant for identifying low nutrition intake
Fish	Included	Relevant for identifying low nutrition intake
Dairy	Included	Relevant for identifying low nutrition intake
Red meat	Included	Relevant for identifying low nutrition intake
environmental_impacts_freshwater_use	Included	Relevant to understanding enviromental footprint on food system components
planetary_impacts_Freshwater	Included	Relevant to understanding food system's impact on planetary boundary value

Describe, Explore and Verify Data Quality

Obesity

Describe Data

Table 2: Data Description of the Variable "Obesity"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	16.31
Obesity	Variance	126.10
	Std. Dev.	11.23
Data Volume (number of observation/rows)	Skewness	0.9259
6460	Kurtosis	4.2428
	Median	16.10
Meaning of the attribute	Mean Abs. Dev.	8.81
	Mode	3.56
Adults aged 18 years and older with a BMI of 30	Minimum	0.37
	Maximum	63.34
Meaning of the attribute in business terms	Range	62.97
	Count	6460
	Sum	105358.26
	1st Quartile	6.71
Attribute types (select from the list)	3rd Quartile	22.55
Continuous	Interquartile Range	15.84
	Missing / Blank	0
****Excel: Summary Statistics ***Stat Tools: One Variable Summary**		

Explore Data

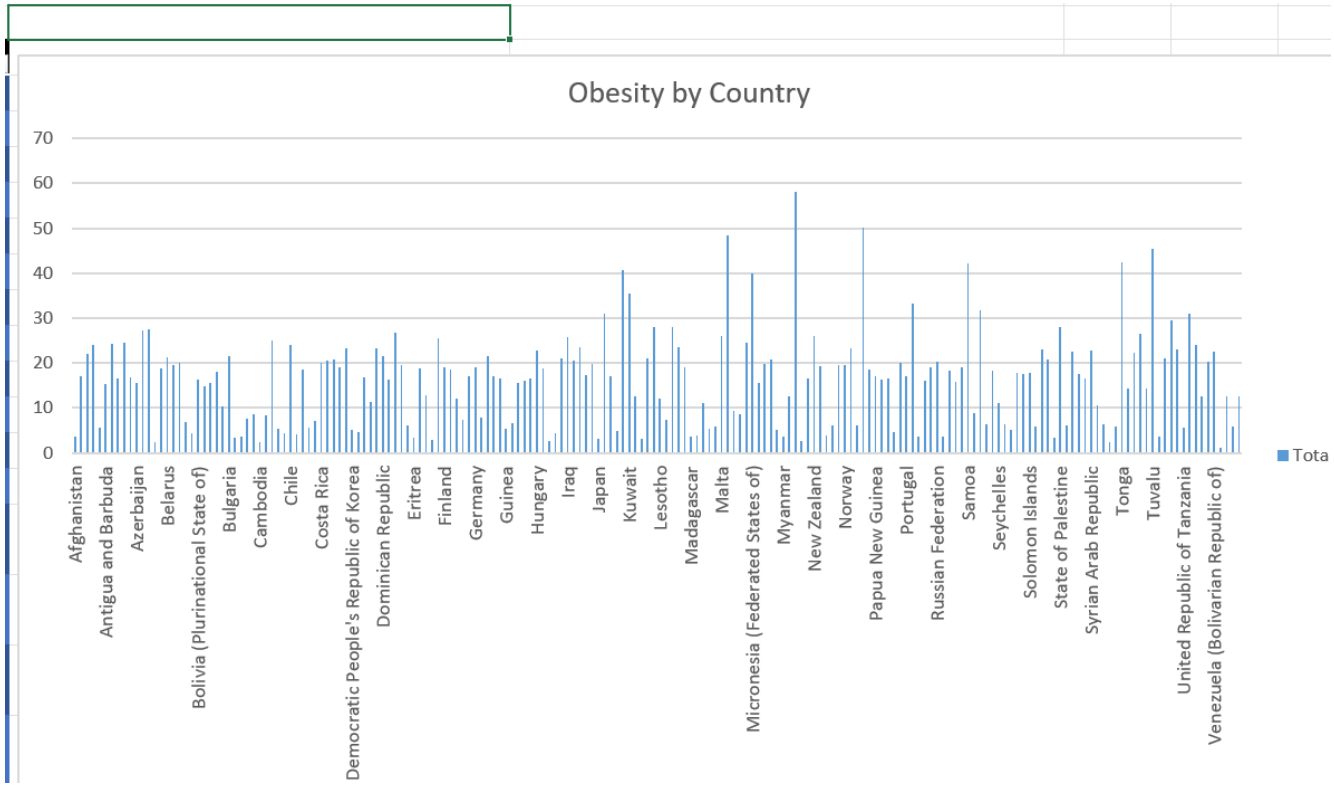


Figure 1: This Bar Graph shows obesity rates per country based on Body Mass Index.

- This bar graph represents the Obesity average by country, calculated by BMI of adults, anything under 18.5 is considered healthy while everything above 25 is overweight.
- There are a few outliers, both too low and high to be correct in regard to BMI average by country.
- African and Asian countries have underweight average compared to all regions.

Verify Data Quality

Table 3: Data Quality Verification of the Variable "Obesity"

Return to List of Topics

					Data Preparation	
Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Country	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	No	n/a		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	lower-case letter, sometimes with an upper-case letter	No	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	n/a		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	Yes	There are a few outliers that skew too low and high in obesity.		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	Yes	Was able to structure the table by making an obesity and year field to organize it.		

Country

Describe Data

Table 4: Data Description of the Variable "Country"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	16.31
Obesity	Variance	126.10
	Std. Dev.	11.23
Data Volume (number of observation/rows)	Skewness	0.9259
6460	Kurtosis	4.2428
	Median	16.10
Meaning of the attribute	Mean Abs. Dev.	8.81
	Mode	3.56
Adults aged 18 years and older with a BMI of 30	Minimum	0.37
	Maximum	63.34
Meaning of the attribute in business terms	Range	62.97
	Count	6460
	Sum	105358.26
	1st Quartile	6.71
Attribute types (select from the list)	3rd Quartile	22.55
Continuous	Interquartile Range	15.84
	Missing / Blank	0

**Excel: Summary Statistics ***Stat Tools: One Variable Summary

Explore Data

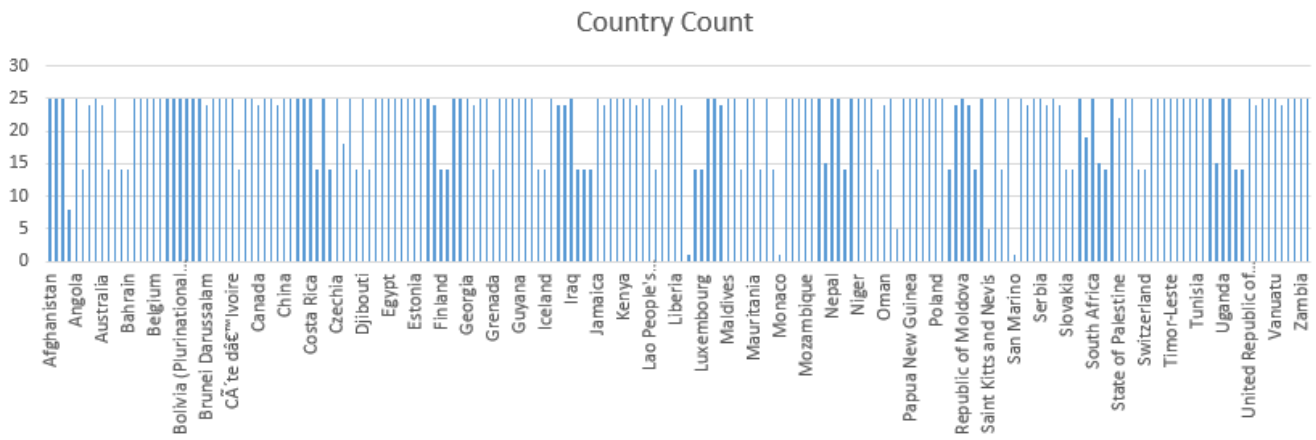


Figure 2: This Bar Graph the data count for each country.

Verify Data Quality

Table 5: Data Quality Verification of the Variable "Country"

Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Country	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	No	n/a		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	lower-case letter, sometimes with an upper-case letter	No	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	n/a		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	Yes	There are a few outliers that skew too low and high in obesity.		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Region

Describe Data

Table 6: Data Description of the Variable "Region"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	16.31
Obesity	Variance	126.10
	Std. Dev.	11.23
Data Volume (number of observation/rows)	Skewness	0.9259
6460	Kurtosis	4.2428
	Median	16.10
Meaning of the attribute	Mean Abs. Dev.	8.81
	Mode	3.56
Adults aged 18 years and older with a BMI of 30	Minimum	0.37
	Maximum	63.34
Meaning of the attribute in business terms	Range	62.97
	Count	6460
	Sum	105358.26
	1st Quartile	6.71
Attribute types (select from the list)	3rd Quartile	22.55
Continuous	Interquartile Range	15.84
	Missing / Blank	0
****Excel: Summary Statistics ***Stat Tools: One Variable Summary**		

Explore Data

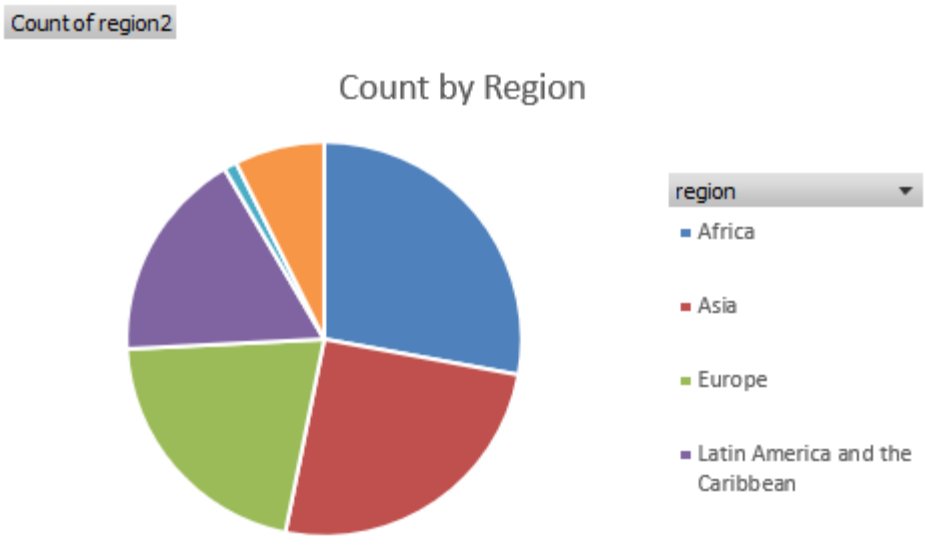


Figure 3: This Pie Chart shows the data count for each region.

Verify Data Quality

Table 7: Data Quality Verification of the Variable "Region"

Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Country	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	No	n/a		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	lower-case letter, sometimes with an upper-case letter	No	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	n/a		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	No	n/a		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Disagg.value

Describe Data

Table 8: Data Description of the Variable "disagg.value"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	16.31
Obesity	Variance	126.10
	Std. Dev.	11.23
Data Volume (number of observation/rows)	Skewness	0.9259
6460	Kurtosis	4.2428
	Median	16.10
Meaning of the attribute	Mean Abs. Dev.	8.81
	Mode	3.56
Adults aged 18 years and older with a BMI of 30	Minimum	0.37
	Maximum	63.34
Meaning of the attribute in business terms	Range	62.97
	Count	6460
	Sum	105358.26
	1st Quartile	6.71
Attribute types (select from the list)	3rd Quartile	22.55
Continuous	Interquartile Range	15.84
	Missing / Blank	0

Explore Data

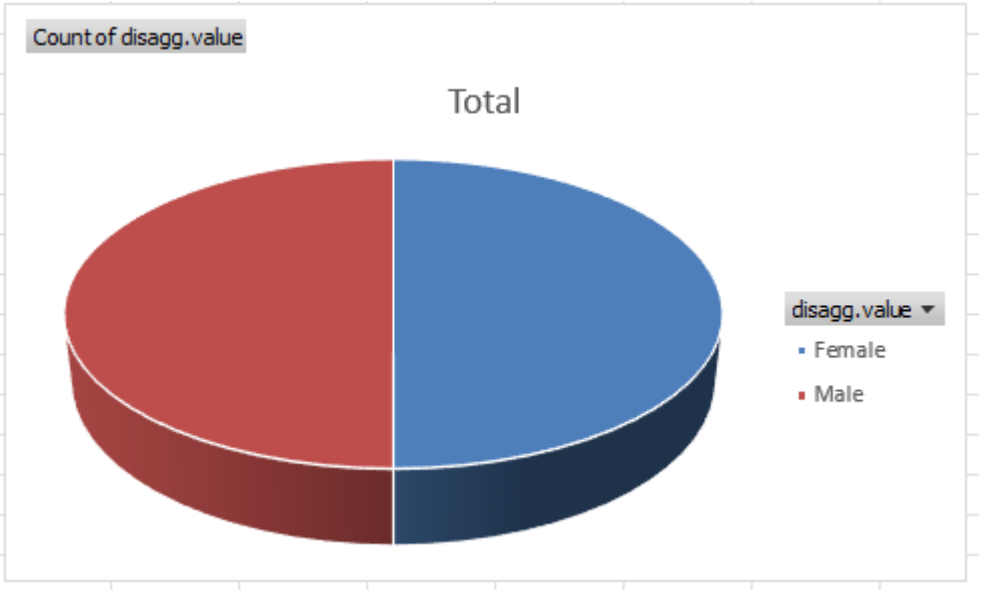


Figure 4: This Pie Chart shows the data count each value in sex per country.

Verify Data Quality

Table 9: Data Quality Verification of the Variable "disagg.value"

Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Country	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	No	n/a		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	lower-case letter, sometimes with an upper-case letter	No	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	n/a		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	No	n/a		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Iso3

Describe Data

Table 10: Data Description of the Variable "disagg.value"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	16.31
Obesity	Variance	126.10
	Std. Dev.	11.23
Data Volume (number of observation/rows)	Skewness	0.9259
6460	Kurtosis	4.2428
	Median	16.10
Meaning of the attribute	Mean Abs. Dev.	8.81
	Mode	3.56
Adults aged 18 years and older with a BMI of 30	Minimum	0.37
	Maximum	63.34
Meaning of the attribute in business terms	Range	62.97
	Count	6460
	Sum	105358.26
	1st Quartile	6.71
Attribute types (select from the list)	3rd Quartile	22.55
Continuous	Interquartile Range	15.84
	Missing / Blank	0

Explore Data

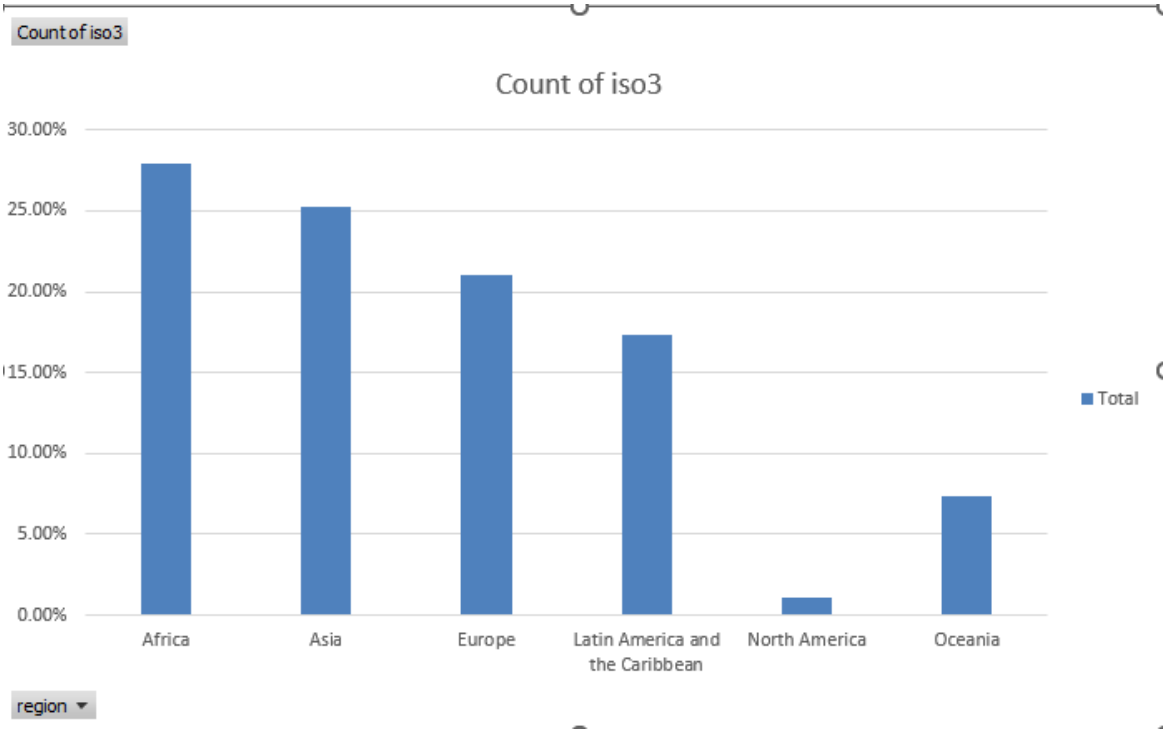


Figure 5: This Bar Graph shows the count for iso3 (ISO code for countries) for each region.

Verify Data Quality

Table 11: Data Quality Verification of the Variable "disagg.value"

Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Country	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	No	n/a		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	lower-case letter, sometimes with an upper-case letter	No	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	n/a		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	No	n/a		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Year

Describe Data

Table 12: Data Description of the Variable "year"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	16.31
Obesity	Variance	126.10
	Std. Dev.	11.23
Data Volume (number of observation/rows)	Skewness	0.9259
6460	Kurtosis	4.2428
	Median	16.10
Meaning of the attribute	Mean Abs. Dev.	8.81
	Mode	3.56
Adults aged 18 years and older with a BMI of 30	Minimum	0.37
	Maximum	63.34
Meaning of the attribute in business terms	Range	62.97
	Count	6460
	Sum	105358.26
	1st Quartile	6.71
Attribute types (select from the list)	3rd Quartile	22.55
Continuous	Interquartile Range	15.84
	Missing / Blank	0

Explore Data

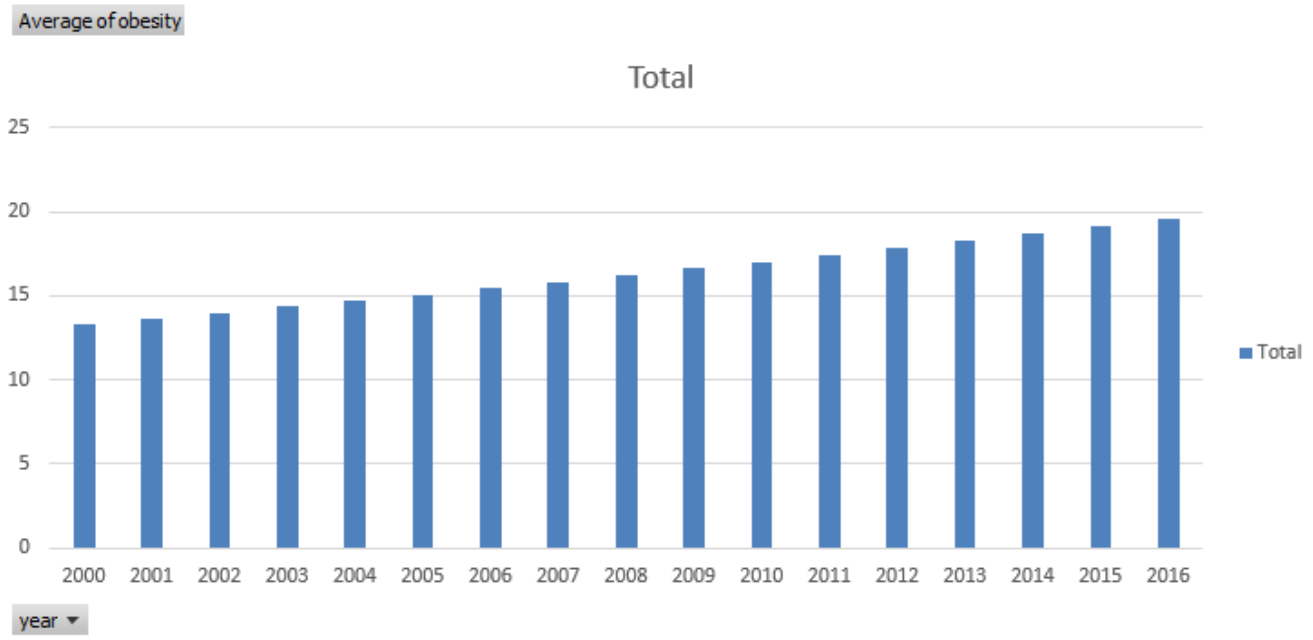


Figure 6: This Bar Graph shows the average rates of Obesity throughout the years.

Verify Data Quality

Table 13: Data Quality Verification of the Variable "year"

Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Country	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	No	n/a		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	lower-case letter, sometimes with an upper-case letter	No	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	n/a		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	No	n/a		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Fruit

Describe Data

Table 14: Data Description of the Variable "Fruit"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	102.7704918
Fruit	Variance	3558.128245
	Std. Dev.	59.65004816
Data Volume (number of observation/rows)	Skewness	2.917105033
183	Kurtosis	17.36024601
	Median	91
Meaning of the attribute	Mean Abs. Dev.	38.28596255
Estimated intake of fruit in adults aged 20 and older (g/day)	Mode	34.8
	Minimum	3.9
	Maximum	499.9
Meaning of the attribute in business terms	Range	496
Estimated intake of fruit in adults aged 20 and older (g/day)	Count	183
	Sum	18807
	1st Quartile	69.4
Attribute types (select from the list)	3rd Quartile	121.5
Continuous	Interquartile Range	52.1
	Missing / Blank	10
****Excel: Summary Statistics ***Stat Tools: One Variable Summary**		

Explore Data

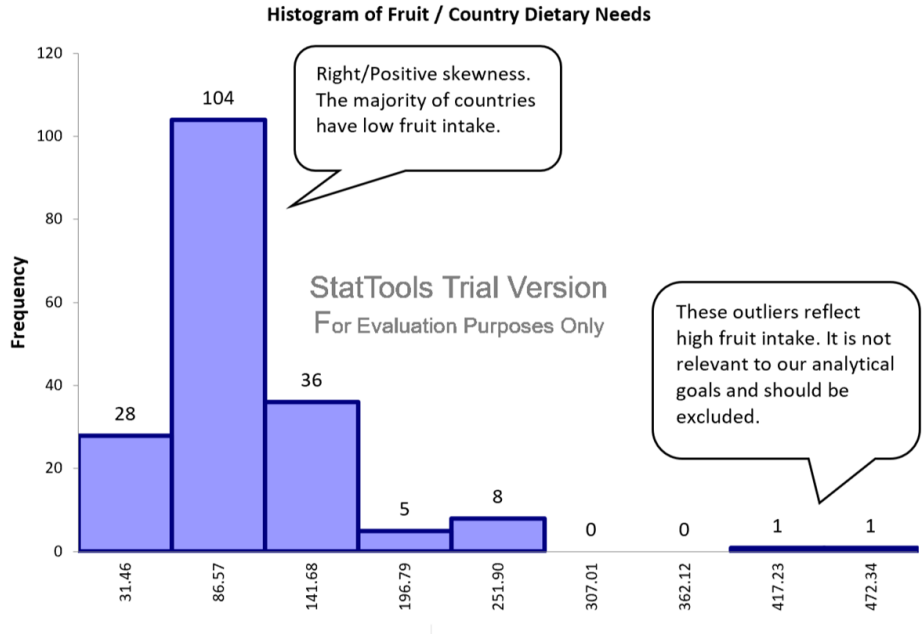


Figure 7: This graph is a histogram for fruit intake to examine country dietary needs.

Verify Data Quality

Table 15: Data Quality Verification of the Variable "Fruit"

	Return to List of Topics				Data Preparation	
Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Fruit	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	Yes	There are 10 missing values		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	N/A	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	Yes	Right-skewness with extremely high outliers, but no noise		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	Yes	High value outliers that are not relevant to countries with low nutritional needs		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Vegetables

Describe Data

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	148.5043716
Vegetables	Variance	5633.715256
	Std. Dev.	75.05807922
Data Volume (number of observation/rows)	Skewness	1.81398324
183	Kurtosis	7.781890431
	Median	131.8
Meaning of the attribute	Mean Abs. Dev.	52.07092478
Estimated intake of vegetables in adults aged 20 and older (g/day)	Mode	69.3
	Minimum	20.2
	Maximum	480.8
Meaning of the attribute in business terms	Range	460.6
Estimated intake of vegetables in adults aged 20 and older (g/day)	Count	183
	Sum	27176.3
	1st Quartile	103.4
Attribute types (select from the list)	3rd Quartile	174.2
Continuous	Interquartile Range	70.8
	Missing / Blank	10
**Excel: Summary Statistics ***Stat Tools: One Variable Summary		

Explore Data

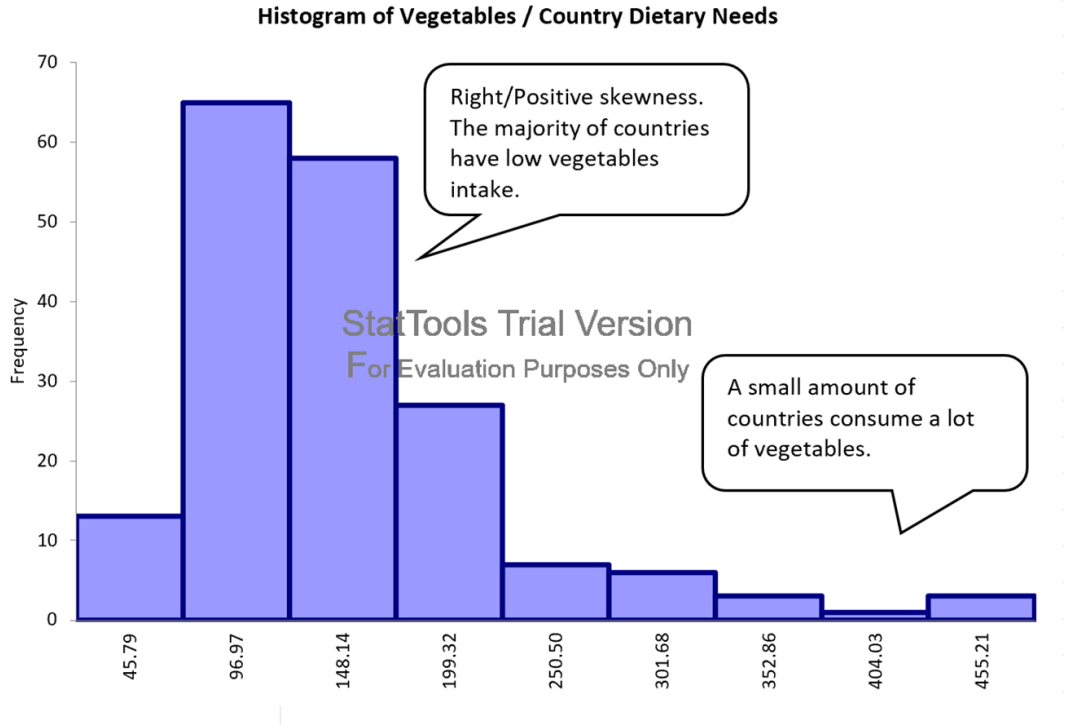


Figure 8: This graph is a histogram for vegetables intake to examine country dietary needs.

Verify Data Quality

Table 16: Data Quality Verification of the Variable "Vegetables"

Return to List of Topics

Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Preparation: Data Cleaning	Data Preparation: Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Vegetables	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	Yes	There are 10 missing values		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	N/A	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	Yes	Right-skewness, but no noise		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	No	n/a		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Legumes

Describe Data

Table 17: Data Description of the Variable "Legumes"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	30.12185792
Legumes	Variance	801.1872119
	Std. Dev.	28.30525061
Data Volume (number of observation/rows)	Skewness	3.619791001
183	Kurtosis	23.05785268
	Median	25.9
Meaning of the attribute	Mean Abs. Dev.	17.22067545
Estimated intake of legumes in adults aged 20 and older (g/day)	Mode	15.2
	Minimum	1
	Maximum	231.1
Meaning of the attribute in business terms	Range	230.1
Estimated intake of legumes in adults aged 20 and older (g/day)	Count	183
	Sum	5512.3
	1st Quartile	13.7
Attribute types (select from the list)	3rd Quartile	37.8
Continuous	Interquartile Range	24.1
	Missing / Blank	10
**Excel: Summary Statistics ***Stat Tools: One Variable Summary		

Explore Data

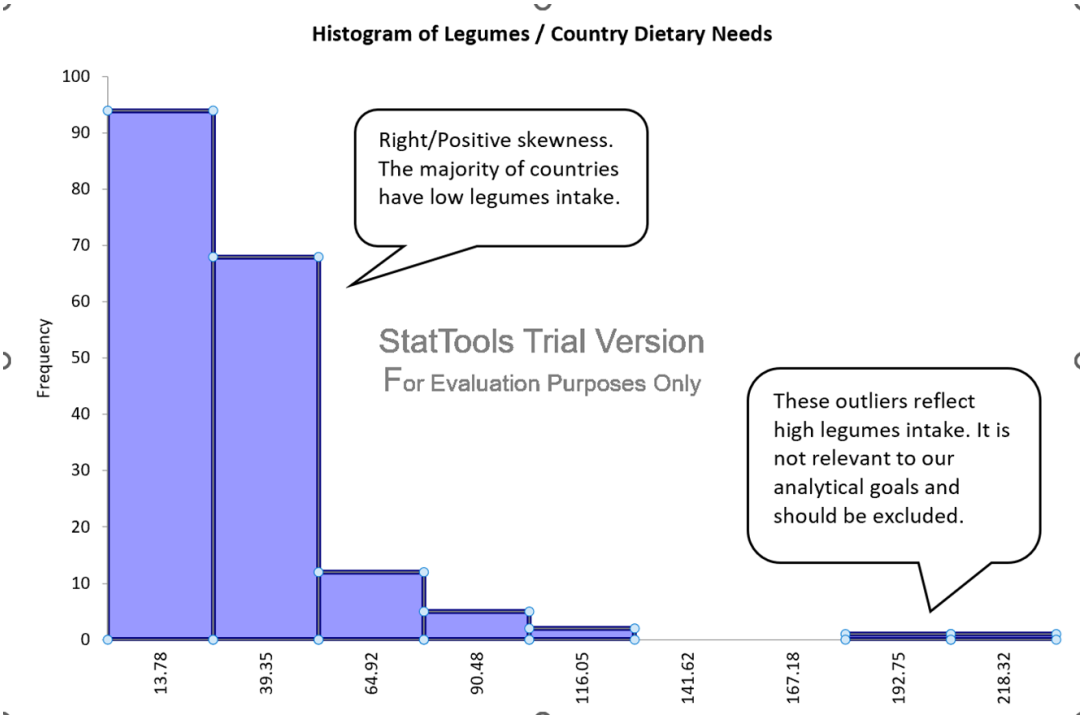


Figure 9: This graph is a histogram for legumes intake to examine country dietary needs.

Verify Data Quality

Table 18: Data Quality Verification of the Variable "Legumes"

	Return to List of Topics				Data Preparation	
Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Legumes	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	Yes	There are 10 missing values		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	N/A	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	Yes	Right-skewness with extremely high outliers, but no noise		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	Yes	High value outliers that are not relevant to countries with low nutritional needs		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Nuts

Describe Data

Table 19: Data Description of the Variable "Nuts"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	9.929508197
Nuts	Variance	110.424619
	Std. Dev.	10.5083119
Data Volume (number of observation/rows)	Skewness	2.67080848
183	Kurtosis	13.47863617
	Median	7.4
Meaning of the attribute	Mean Abs. Dev.	7.063835886
Estimated intake of nuts in adults aged 20 and older (g/day)	Mode	3.2
	Minimum	0
	Maximum	75
Meaning of the attribute in business terms	Range	75
Estimated intake of nuts in adults aged 20 and older (g/day)	Count	183
	Sum	1817.1
	1st Quartile	3.1
Attribute types (select from the list)	3rd Quartile	12.6
Continuous	Interquartile Range	9.5
	Missing / Blank	10
**Excel: Summary Statistics ***Stat Tools: One Variable Summary		

Explore Data

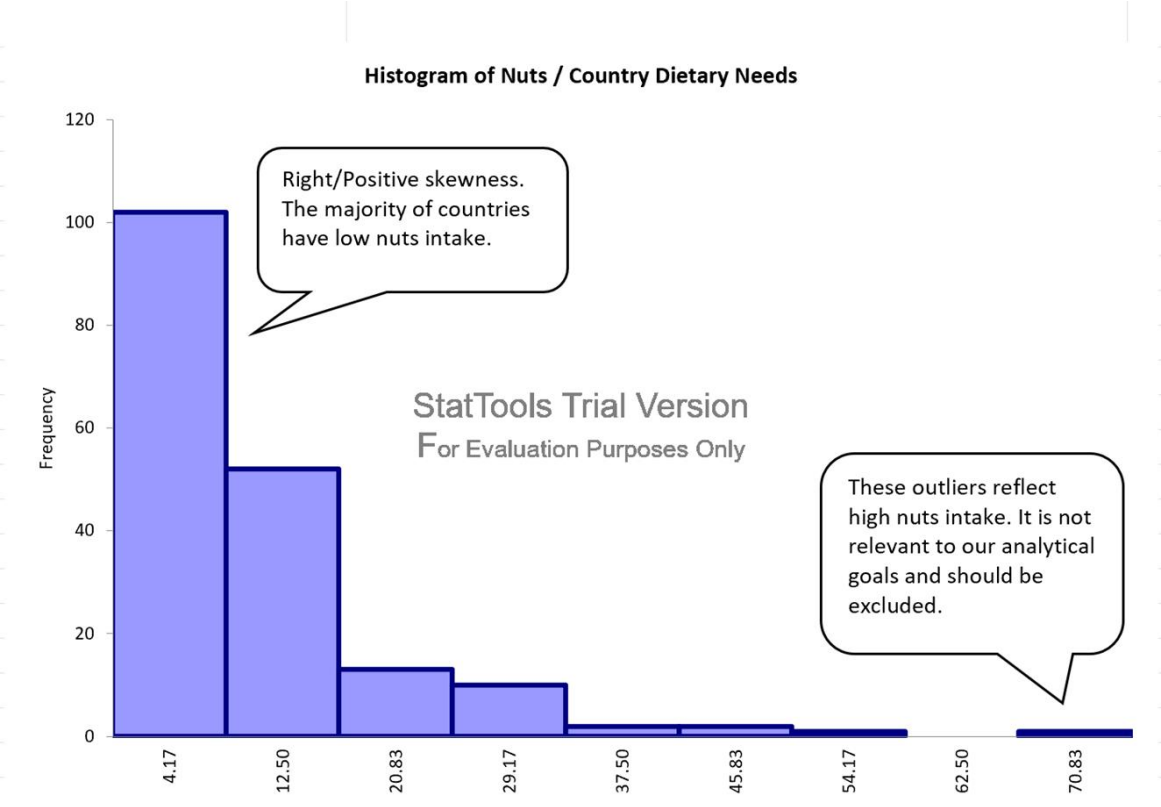


Figure 10: This graph is a histogram for nuts intake to examine country dietary needs.

Verify Data Quality

Table 20: Data Quality Verification of the Variable "Nuts"

	Return to List of Topics				Data Preparation	
Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
Nuts	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	n/a		
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	Yes	There are 10 missing values		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	N/A	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	Yes	Right-skewness with extremely high outliers, but no noise		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	Yes	High value outliers that are not relevant to countries with low nutritional needs		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Whole grains

Describe Data

Table 21: Data Description of the Variable "Whole grains"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	42.82568306
Whole grains	Variance	1125.506095
	Std. Dev.	33.54856323
Data Volume (number of observation/rows)	Skewness	2.000310021
183	Kurtosis	8.454601651
	Median	34.5
Meaning of the attribute	Mean Abs. Dev.	23.90575413
Estimated intake of whole grains in adults aged 20 and older (g/day)	Mode	25
	Minimum	0.1
	Maximum	194.8
Meaning of the attribute in business terms	Range	194.7
Estimated intake of whole grains in adults aged 20 and older (g/day)	Count	183
	Sum	7837.1
	1st Quartile	22
Attribute types (select from the list)	3rd Quartile	57.2
Continuous	Interquartile Range	35.2
	Missing / Blank	10
**Excel: Summary Statistics ***Stat Tools: One Variable Summary		

Explore Data

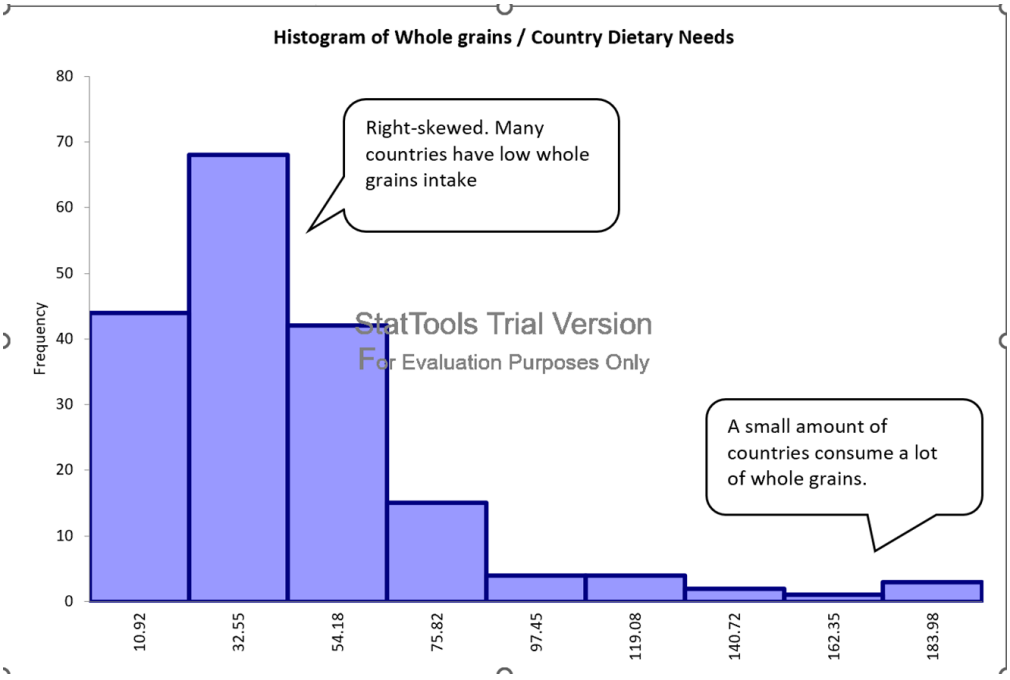


Figure 11: This graph is a histogram for whole grains intake to examine country dietary needs.

Verify Data Quality

Table 22: Data Quality Verification of the Variable "Whole grains"

	Return to List of Topics				Data Preparation	
Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Whole grains	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	Yes			
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	N/A	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	Yes	Right-skewness, but no noise		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	No	n/a		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Fish

Describe Data

Table 23: Data Description of the Variable "Fish"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	34.06120219
Fish	Variance	539.5613985
	Std. Dev.	23.22846096
Data Volume (number of observation/rows)	Skewness	2.127054522
183	Kurtosis	11.82785391
	Median	30.6
Meaning of the attribute	Mean Abs. Dev.	16.46452268
Estimated intake of fish in adults aged 20 and older (g/day)	Mode	11.9
	Minimum	0
	Maximum	174.7
Meaning of the attribute in business terms	Range	174.7
Estimated intake of fish in adults aged 20 and older (g/day)	Count	183
	Sum	6233.2
	1st Quartile	18.1
Attribute types (select from the list)	3rd Quartile	44.4
Continuous	Interquartile Range	26.3
	Missing / Blank	10
**Excel: Summary Statistics ***Stat Tools: One Variable Summary		

Explore Data

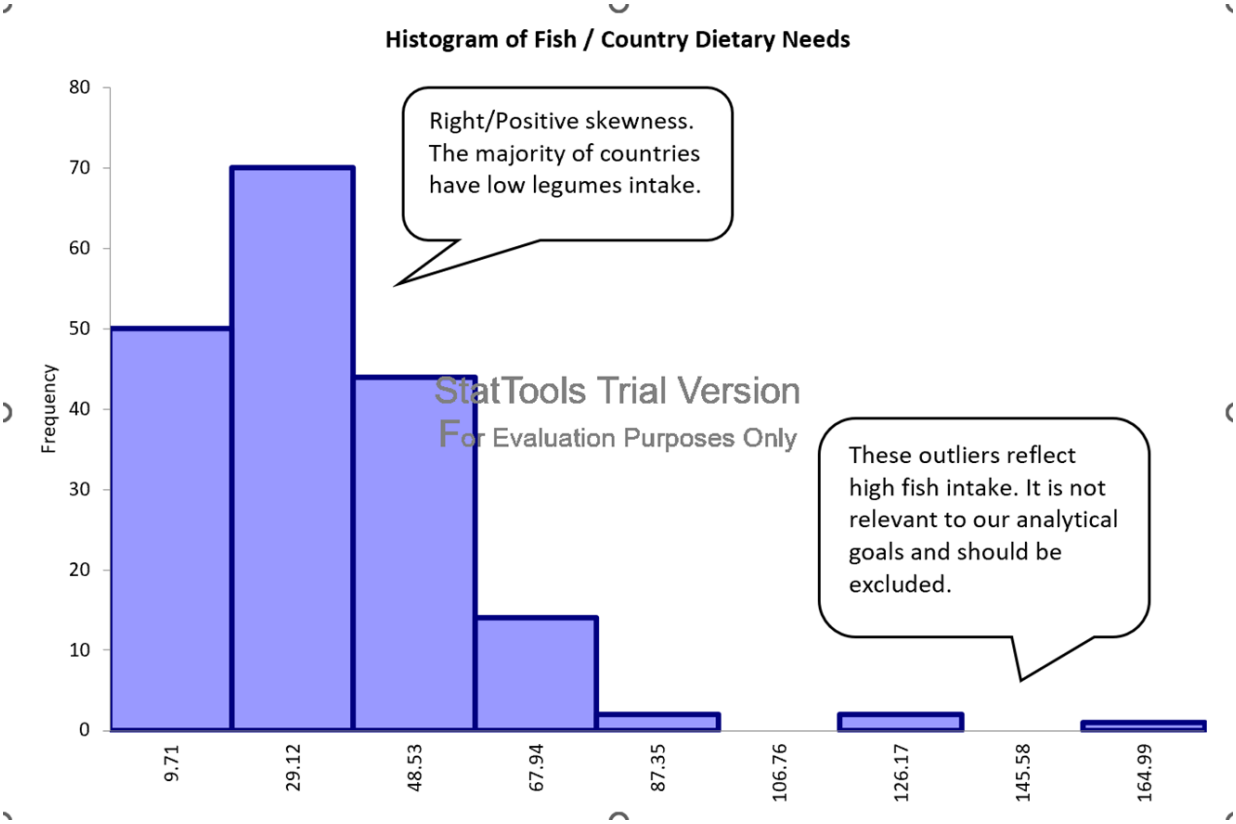


Figure 12: This graph is a histogram for fish intake to examine country dietary needs.

Verify Data Quality

Table 24: Data Quality Verification of the Variable "Fish"

Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Fish	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	Yes	There are 10 missing values		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	N/A	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	Yes	Right-skewness with extremely high outliers, but no noise		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	Yes	High value outliers that are not relevant to countries with low nutritional needs		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Dairy

Describe Data

Table 25: Data Description of the Variable "Dairy"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	309.3628415
Dairy	Variance	79584.07048
	Std. Dev.	282.1064878
Data Volume (number of observation/rows)	Skewness	1.393419232
183	Kurtosis	5.257315347
	Median	204.6
Meaning of the attribute	Mean Abs. Dev.	226.1979516
Estimated intake of dairy in adults aged 20 and older (g/day)	Mode	54.1
	Minimum	23.4
	Maximum	1557
Meaning of the attribute in business terms	Range	1533.6
Estimated intake of dairy in adults aged 20 and older (g/day)	Count	183
	Sum	56613.4
	1st Quartile	80.3
Attribute types (select from the list)	3rd Quartile	481.1
Continuous	Interquartile Range	400.8
	Missing / Blank	10
**Excel: Summary Statistics ***Stat Tools: One Variable Summary		

Explore Data

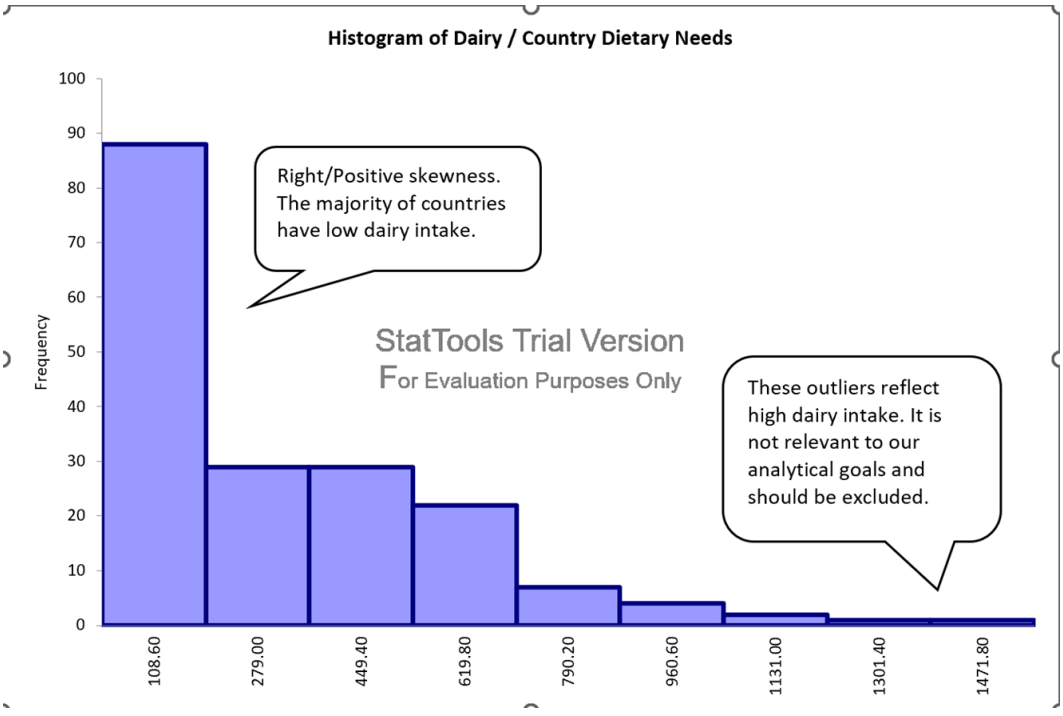


Figure 13: This graph is a histogram for dairy intake to examine country dietary needs.

Verify Data Quality

Table 26: Data Quality Verification of the Variable "Dairy"

	Return to List of Topics				Data Preparation	
Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Dairy	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	Yes	There are 10 missing values		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	N/A	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	Yes	Right-skewness with extremely high outliers, but no noise		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	Yes	High value outliers that are not relevant to countries with low nutritional needs		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Red meat

Describe Data

Table 27: Data Description of the Variable "Red meat"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	51.46557377
Red meat	Variance	1775.430622
	Std. Dev.	42.13585909
Data Volume (number of observation/rows)	Skewness	2.093689835
183	Kurtosis	8.961554909
	Median	41.7
Meaning of the attribute	Mean Abs. Dev.	29.87845561
Estimated intake of red meat in adults aged 20 and older (g/day)	Mode	27
	Minimum	2.7
	Maximum	251.5
Meaning of the attribute in business terms	Range	248.8
Estimated intake of red meat in adults aged 20 and older (g/day)	Count	183
	Sum	9418.2
	1st Quartile	22.3
Attribute types (select from the list)	3rd Quartile	66.5
Continuous	Interquartile Range	44.2
	Missing / Blank	10
**Excel: Summary Statistics ***Stat Tools: One Variable Summary		

Explore Data

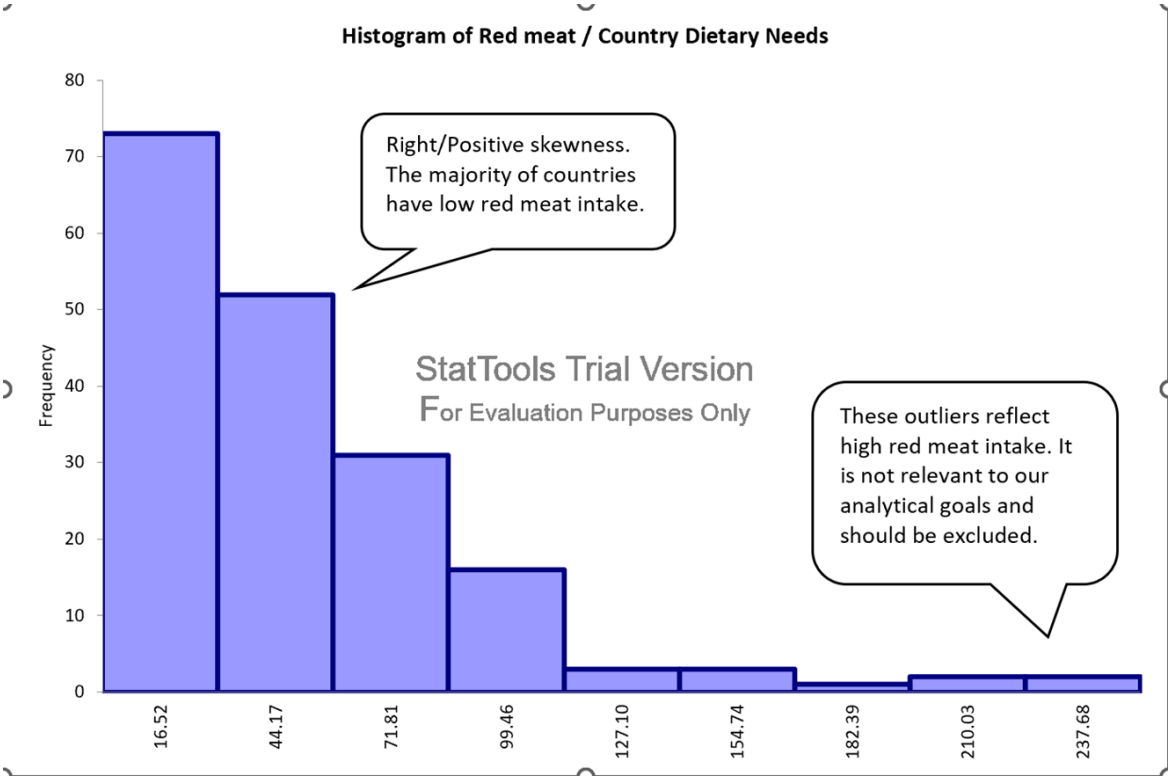


Figure 14: This graph is a histogram for red meat intake to examine country dietary needs.

Verify Data Quality

Table 28: Data Quality Verification of the Variable "Red meat"

Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	n/a		
Red meat	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	n/a		
	Missing Attribute or blank fields	How will you address this?	Yes	There are 10 missing values		
	Duplicate	Duplicated records (observations)	No	n/a		
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	N/A	n/a		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	Yes	Right-skewness with extremely high outliers, but no noise		
	Plausibility	E.g., all fields having the same or nearly the same values	No	n/a		
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	n/a		
	High Cardinality	A high number of values in a set	No	n/a		
	Outliers	An observation that lies well outside of the norm.	Yes	High value outliers that are not relevant to countries with low nutritional needs		
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	n/a		
	Sparseness	Any data which as very large zero value and very little no zero value	No	n/a		
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	n/a		
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	n/a		

Environmental_impacts_freshwater_use

Describe Data

Table 29: Describe Description of variable "Environmental_impacts_freshwater_use"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	1.73
environmental_impacts_freshwater_use	Variance	221.35
	Std. Dev.	14.88
Data Volume (number of observation/rows)	Skewness	25.3513
2156	Kurtosis	813.7650
	Median	0.05
Meaning of the attribute	Mean Abs. Dev.	2.77
Environmental footprint of food system components on freshwater use	Mode	0.00
	Minimum	0.00
	Maximum	530.35
Meaning of the attribute in business terms	Range	530.35
	Count	2071
	Sum	3592.29
	1st Quartile	0.01
Attribute types (select from the list)	3rd Quartile	0.36
Continuous	Interquartile Range	0.36
	Missing / Blank	85
**Excel: Summary Statistics ***Stat Tools: One Variable Summary		

Explore Data

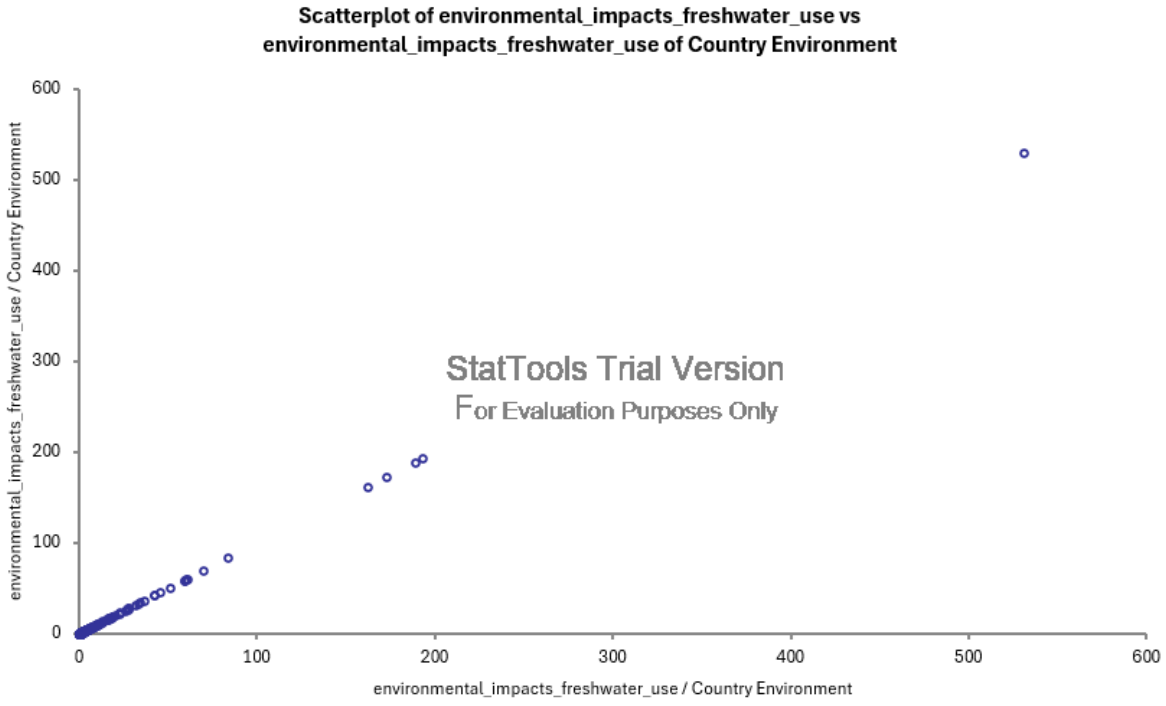


Figure 15: This scatterplot chart reveals any high impact on freshwater from producing certain foods.

Verify Data Quality

Table 30: Data Quality Verification of the variable "Environmental_impacts_freshwater_use"

Return to List of Topics

Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment
environmental_impacts_freshwater_use	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	Yes	
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	Yes	
	Missing Attribute or blank fields	How will you address this?	Yes	I will make a close assessment on the blank fields and see which best value fits or whether it needs to be ommitted from the dataset.
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	lower-case letter, sometimes with an upper-case letter	No	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	Yes	See whether the outlier is important to the dataset. If not, might need removal for more accurate analysis.
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which as very large zero value and very little no zero value	No	
	Irrelevant Input	Does not provide information about the target (dependent Variable)	N/A	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

Planetary_impacts_Freshwater

Describe Data

Table 31: Describe Description of variable "Planetary_impacts_Freshwater"

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	18.48
planetary_impacts_Freshwater	Variance	1188.07
	Std. Dev.	34.47
Data Volume (number of observation/rows)	Skewness	2.9879
2156	Kurtosis	12.2977
	Median	4.01
Meaning of the attribute	Mean Abs. Dev.	21.65
Food system impact on planetary boundary value (%) of freshwater	Mode	0.07
	Minimum	0.00
	Maximum	238.25
Meaning of the attribute in business terms	Range	238.25
	Count	2155
	Sum	39819.72
	1st Quartile	1.30
Attribute types (select from the list)	3rd Quartile	19.27
Continuous	Interquartile Range	17.96
	Missing / Blank	1

Explore Data

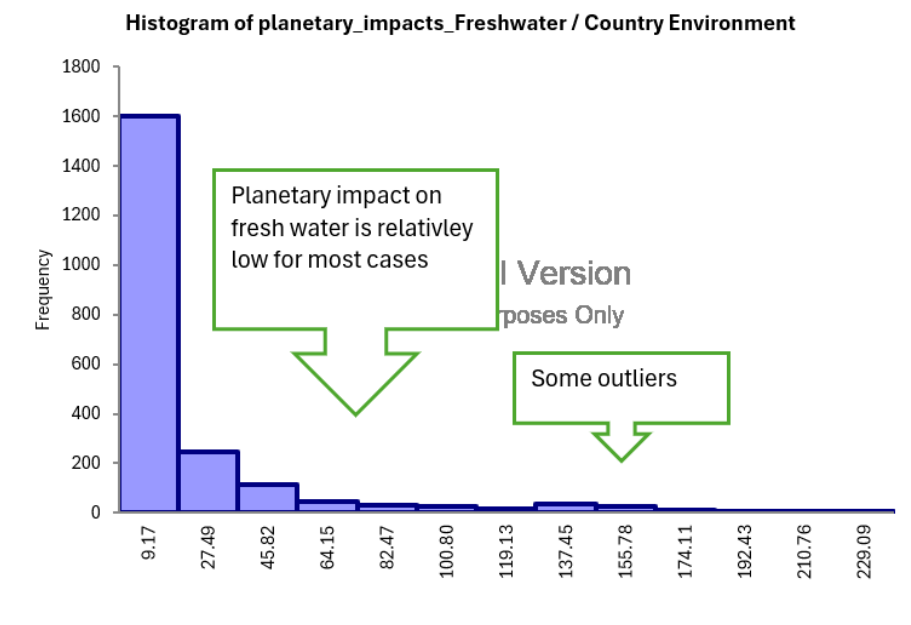


Figure 16: This histogram chart reveals any planetary impacts to freshwater while producing foods

Verify Data Quality

Table 32: Data Quality Verification of the variable "Planetary_impacts_Freshwater"

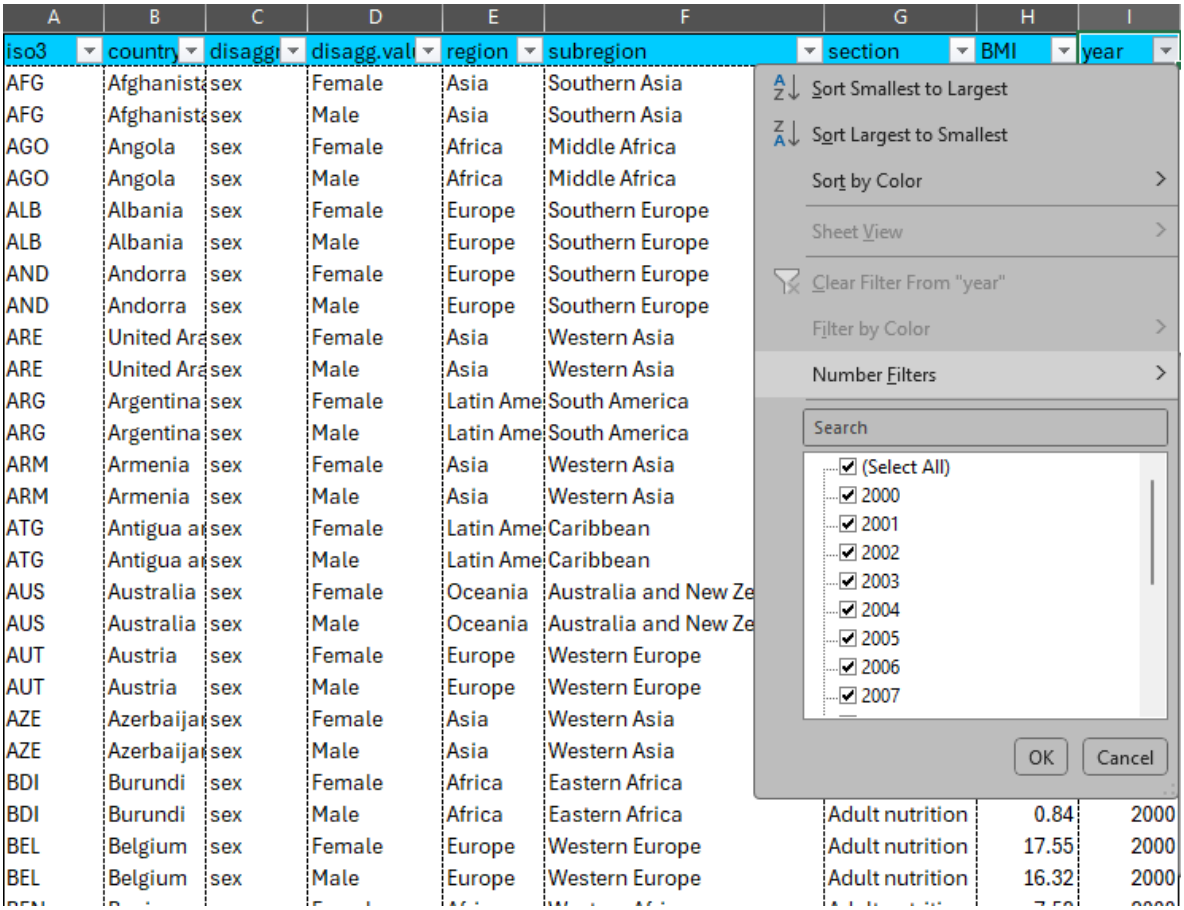
Variable Name	Data Quality Issue	Description/ Example	Problem (Select from the list)	Assessment
Planetary_impacts_Freshwater	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blank fields	How will you address this?	Yes	Just one missing value will evaluate the best fitting solution
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	lower-case letter, sometimes with an upper-case letter	No	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	Yes	See whether the outlier is important to the dataset. If not, might need removal for more accurate analysis.
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which as very large zero value and very little no zero value	No	
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

Data Preparation

Clean Data

- Restructuring the new Country Adult BMI table by reorganizing only the data needed.
- The raw dataset contained one column per year where all BMI values were under, after I made a year column where I added all values under and then BMI in separate column where it is directly connected to year.

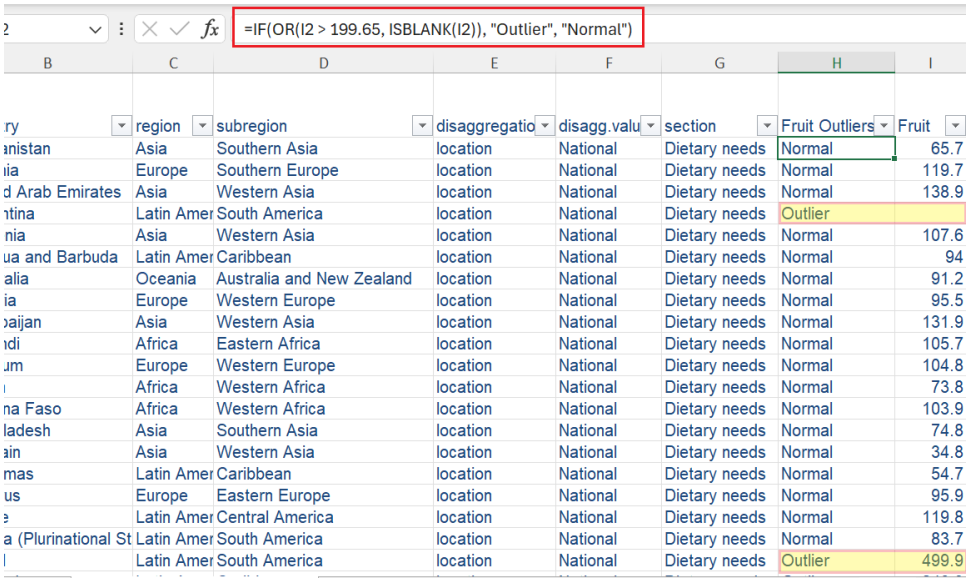
Table 33: Table showing all data included in the restructured table, with BMI and Year sorted vertically in their respective fields.



A	B	C	D	E	F	G	H	I
iso3	country	disaggr	disagg.valu	region	subregion	section	BMI	year
AFG	Afghanista	sex	Female	Asia	Southern Asia			
AFG	Afghanista	sex	Male	Asia	Southern Asia			
AGO	Angola	sex	Female	Africa	Middle Africa			
AGO	Angola	sex	Male	Africa	Middle Africa			
ALB	Albania	sex	Female	Europe	Southern Europe			
ALB	Albania	sex	Male	Europe	Southern Europe			
AND	Andorra	sex	Female	Europe	Southern Europe			
AND	Andorra	sex	Male	Europe	Southern Europe			
ARE	United Ara	sex	Female	Asia	Western Asia			
ARE	United Ara	sex	Male	Asia	Western Asia			
ARG	Argentina	sex	Female	Latin Ame	South America			
ARG	Argentina	sex	Male	Latin Ame	South America			
ARM	Armenia	sex	Female	Asia	Western Asia			
ARM	Armenia	sex	Male	Asia	Western Asia			
ATG	Antigua ar	sex	Female	Latin Ame	Caribbean			
ATG	Antigua ar	sex	Male	Latin Ame	Caribbean			
AUS	Australia	sex	Female	Oceania	Australia and New Ze			
AUS	Australia	sex	Male	Oceania	Australia and New Ze			
AUT	Austria	sex	Female	Europe	Western Europe			
AUT	Austria	sex	Male	Europe	Western Europe			
AZE	Azerbaijar	sex	Female	Asia	Western Asia			
AZE	Azerbaijar	sex	Male	Asia	Western Asia			
BDI	Burundi	sex	Female	Africa	Eastern Africa			
BDI	Burundi	sex	Male	Africa	Eastern Africa	Adult nutrition	0.84	2000
BEL	Belgium	sex	Female	Europe	Western Europe	Adult nutrition	17.55	2000
BEL	Belgium	sex	Male	Europe	Western Europe	Adult nutrition	16.32	2000

- To determine outliers, the Interquartile Rule was used.
- Cleaning Country Dietary Needs Fruit variable:
    - Low outliers: Q1 (69.4) - IQR (52.1) x 1.5 = -8.75. Since the low outlier threshold is below the minimum value, it can be ignored.
    - High outliers: Q3 (121.5) + IQR (52.1) x 1.5 = 199.65
    - Utilize IF statement to determine outliers. The statement will return "Outlier" if the value is greater than the higher outlier threshold.

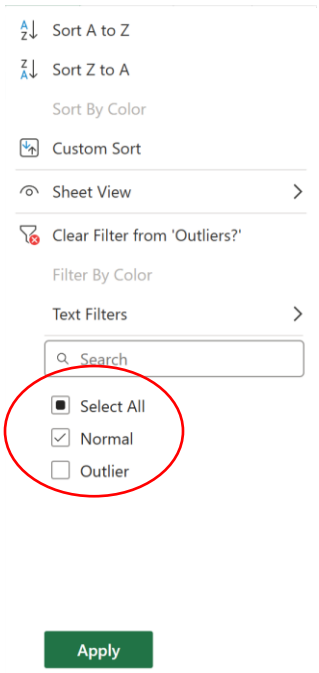
Table 34: Using IF statement to determine outliers for Fruit variable. Missing values will also be categorized as outliers in order to obtain a filtered average to clean missing data.



B	C	D	E	F	G	H	I
ry	region	subregion	disaggregatio	disagg.valu	section	Fruit Outliers	Fruit
anistan	Asia	Southern Asia	location	National	Dietary needs	Normal	65.7
ia	Europe	Southern Europe	location	National	Dietary needs	Normal	119.7
d Arab Emirates	Asia	Western Asia	location	National	Dietary needs	Normal	138.9
ntina	Latin Amer	South America	location	National	Dietary needs	Outlier	
nia	Asia	Western Asia	location	National	Dietary needs	Normal	107.6
ua and Barbuda	Latin Amer	Caribbean	location	National	Dietary needs	Normal	94
alia	Oceania	Australia and New Zealand	location	National	Dietary needs	Normal	91.2
ia	Europe	Western Europe	location	National	Dietary needs	Normal	95.5
baijan	Asia	Western Asia	location	National	Dietary needs	Normal	131.9
ndi	Africa	Eastern Africa	location	National	Dietary needs	Normal	105.7
um	Europe	Western Europe	location	National	Dietary needs	Normal	104.8
	Africa	Western Africa	location	National	Dietary needs	Normal	73.8
na Faso	Africa	Western Africa	location	National	Dietary needs	Normal	103.9
ladesh	Asia	Southern Asia	location	National	Dietary needs	Normal	74.8
ain	Asia	Western Asia	location	National	Dietary needs	Normal	34.8
mas	Latin Amer	Caribbean	location	National	Dietary needs	Normal	54.7
us	Europe	Eastern Europe	location	National	Dietary needs	Normal	95.9
e	Latin Amer	Central America	location	National	Dietary needs	Normal	119.8
a (Plurinational St	Latin Amer	South America	location	National	Dietary needs	Normal	83.7
	Latin Amer	South America	location	National	Dietary needs	Outlier	499.9

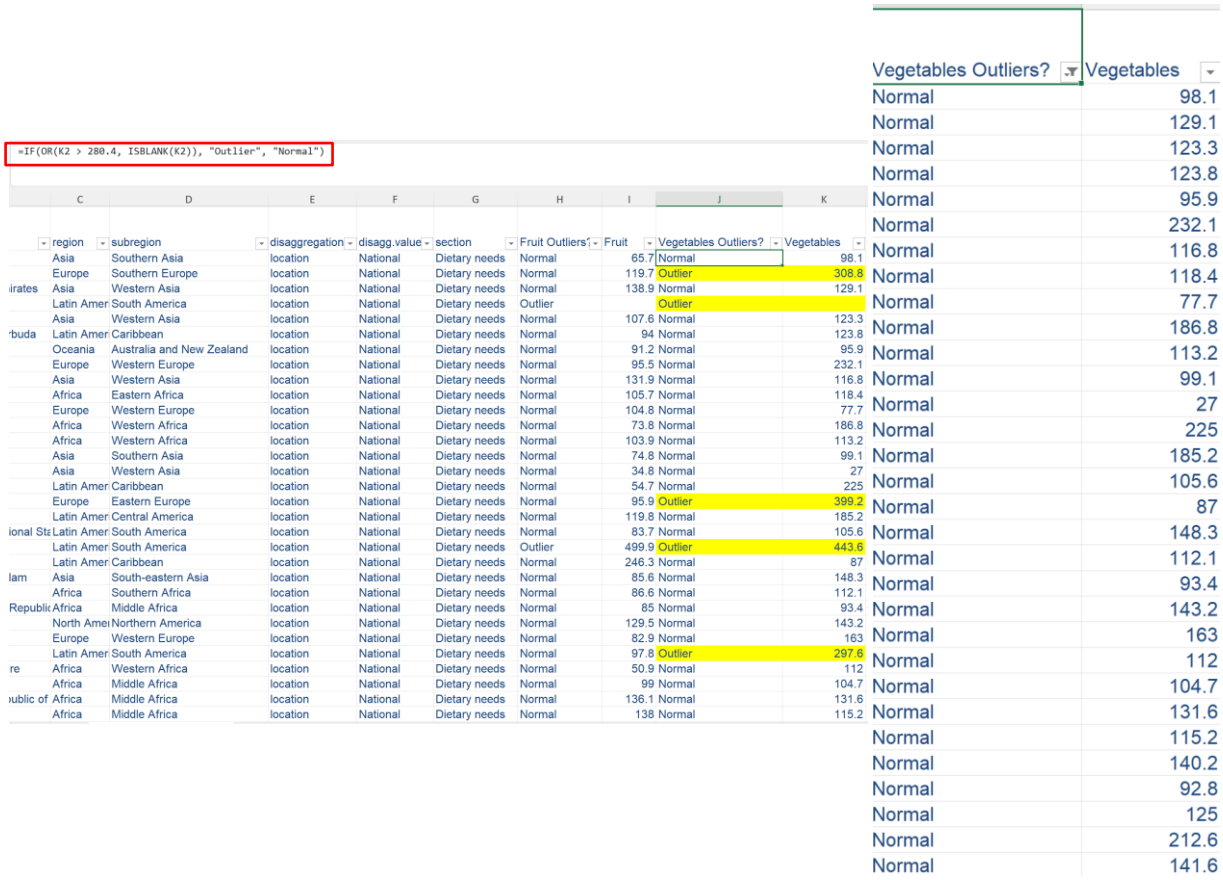
- Filter out outliers by deselecting “Outlier”

Table 35: Using filters to filter out outliers



- Cleaning Country Dietary Needs Vegetables variable:
  - Low outliers: Q1 (103.4) - IQR (70.8) x 1.5 = -2.8
  - High outliers: Q3 (174.2) + IQR (70.8) x 1.5 = 280.4
  - Repeat steps from Fruit variable:

Table 36: Using IF statement to determine outliers for Vegetables variable. Result of filtered outliers using excel filter on Vegetables Outliers.



```
=IF(OR(K2 > 280.4, ISBLANK(K2)), "Outlier", "Normal")
```

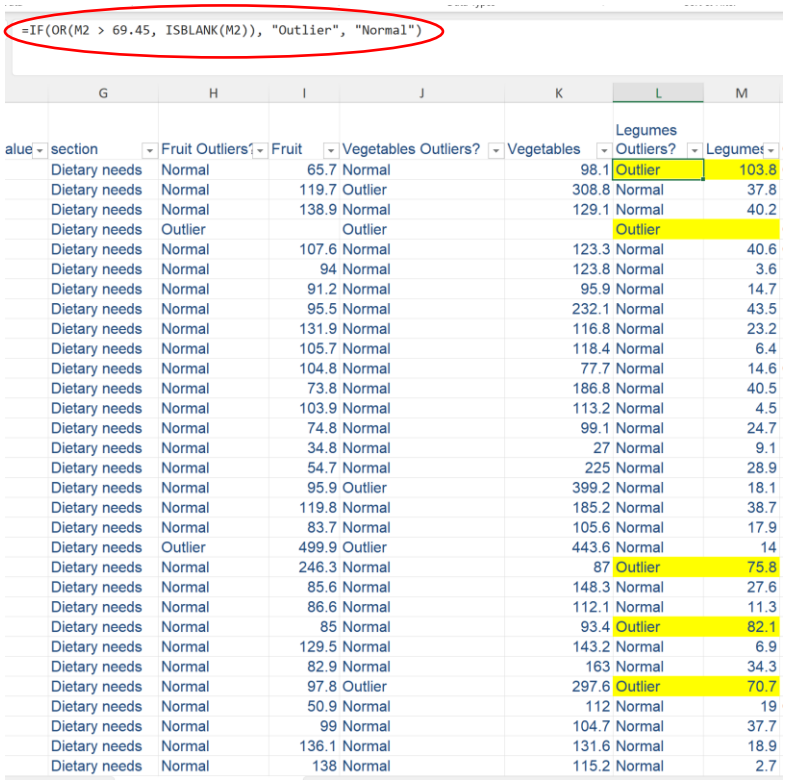
	region	subregion	disaggregation	disagg.value	section	Fruit Outliers	Fruit	Vegetables Outliers?	Vegetables
	Asia	Southern Asia	location	National	Dietary needs	Normal	65.7	Normal	98.1
	Europe	Southern Europe	location	National	Dietary needs	Normal	119.7	Outlier	308.8
irates	Asia	Western Asia	location	National	Dietary needs	Normal	138.9	Normal	129.1
	Latin Amer	South America	location	National	Dietary needs	Outlier		Outlier	
	Asia	Western Asia	location	National	Dietary needs	Normal	107.6	Normal	123.3
buda	Latin Amer	Caribbean	location	National	Dietary needs	Normal	94	Normal	123.8
	Oceania	Australia and New Zealand	location	National	Dietary needs	Normal	91.2	Normal	95.9
	Europe	Western Europe	location	National	Dietary needs	Normal	95.5	Normal	232.1
	Asia	Western Asia	location	National	Dietary needs	Normal	131.9	Normal	116.8
	Africa	Eastern Africa	location	National	Dietary needs	Normal	105.7	Normal	118.4
	Europe	Western Europe	location	National	Dietary needs	Normal	104.8	Normal	77.7
	Africa	Western Africa	location	National	Dietary needs	Normal	73.8	Normal	186.8
	Africa	Western Africa	location	National	Dietary needs	Normal	103.9	Normal	113.2
	Asia	Southern Asia	location	National	Dietary needs	Normal	74.8	Normal	99.1
	Asia	Western Asia	location	National	Dietary needs	Normal	34.8	Normal	27
	Latin Amer	Caribbean	location	National	Dietary needs	Normal	54.7	Normal	225
	Europe	Eastern Europe	location	National	Dietary needs	Normal	95.9	Outlier	399.2
	Latin Amer	Central America	location	National	Dietary needs	Normal	119.8	Normal	185.2
ional Sta	Latin Amer	South America	location	National	Dietary needs	Normal	83.7	Normal	105.6
	Latin Amer	South America	location	National	Dietary needs	Outlier	499.9	Outlier	443.6
	Latin Amer	Caribbean	location	National	Dietary needs	Normal	246.3	Normal	87
lam	Asia	South-eastern Asia	location	National	Dietary needs	Normal	85.6	Normal	148.3
	Africa	Southern Africa	location	National	Dietary needs	Normal	86.6	Normal	112.1
Republic	Africa	Middle Africa	location	National	Dietary needs	Normal	85	Normal	93.4
	North Amer	Northern America	location	National	Dietary needs	Normal	129.5	Normal	143.2
	Europe	Western Europe	location	National	Dietary needs	Normal	82.9	Normal	163
	Latin Amer	South America	location	National	Dietary needs	Normal	97.8	Outlier	297.6
re	Africa	Western Africa	location	National	Dietary needs	Normal	50.9	Normal	112
	Africa	Middle Africa	location	National	Dietary needs	Normal	99	Normal	104.7
ublic of	Africa	Middle Africa	location	National	Dietary needs	Normal	136.1	Normal	131.6
	Africa	Middle Africa	location	National	Dietary needs	Normal	138	Normal	115.2

Vegetables Outliers?	Vegetables
Normal	98.1
Normal	129.1
Normal	123.3
Normal	123.8
Normal	95.9
Normal	232.1
Normal	116.8
Normal	118.4
Normal	77.7
Normal	186.8
Normal	113.2
Normal	99.1
Normal	27
Normal	225
Normal	185.2
Normal	105.6
Normal	87
Normal	148.3
Normal	112.1
Normal	93.4
Normal	143.2
Normal	163
Normal	112
Normal	104.7
Normal	131.6
Normal	115.2
Normal	140.2
Normal	92.8
Normal	125
Normal	212.6
Normal	141.6

- Cleaning Country Dietary Needs Legumes variable:
  - Low outliers: Q1 (13.7) - IQR (21.1) x 1.5 = -17.95
  - High outliers: Q3 (37.8) + IQR (21.1) x 1.5 = 69.45

- Repeat steps from Fruit variable:

Table 37: Using IF statement to determine outliers for Legumes variable. Result of filtered outliers using excel filter on Legumes Outliers.



```
=IF(OR(M2 > 69.45, ISBLANK(M2)), "Outlier", "Normal")
```

alue	section	Fruit Outliers?	Fruit	Vegetables Outliers?	Vegetables	Legumes Outliers?	Legumes
	Dietary needs	Normal	65.7	Normal	98.1	Outlier	103.8
	Dietary needs	Normal	119.7	Outlier	308.8	Normal	37.8
	Dietary needs	Normal	138.9	Normal	129.1	Normal	40.2
	Dietary needs	Outlier		Outlier		Outlier	
	Dietary needs	Normal	107.6	Normal	123.3	Normal	40.6
	Dietary needs	Normal	94	Normal	123.8	Normal	3.6
	Dietary needs	Normal	91.2	Normal	95.9	Normal	14.7
	Dietary needs	Normal	95.5	Normal	232.1	Normal	43.5
	Dietary needs	Normal	131.9	Normal	116.8	Normal	23.2
	Dietary needs	Normal	105.7	Normal	118.4	Normal	6.4
	Dietary needs	Normal	104.8	Normal	77.7	Normal	14.6
	Dietary needs	Normal	73.8	Normal	186.8	Normal	40.5
	Dietary needs	Normal	103.9	Normal	113.2	Normal	4.5
	Dietary needs	Normal	74.8	Normal	99.1	Normal	24.7
	Dietary needs	Normal	34.8	Normal	27	Normal	9.1
	Dietary needs	Normal	54.7	Normal	225	Normal	28.9
	Dietary needs	Normal	95.9	Outlier	399.2	Normal	18.1
	Dietary needs	Normal	119.8	Normal	185.2	Normal	38.7
	Dietary needs	Normal	83.7	Normal	105.6	Normal	17.9
	Dietary needs	Outlier	499.9	Outlier	443.6	Normal	14
	Dietary needs	Normal	246.3	Normal	87	Outlier	75.8
	Dietary needs	Normal	85.6	Normal	148.3	Normal	27.6
	Dietary needs	Normal	86.6	Normal	112.1	Normal	11.3
	Dietary needs	Normal	85	Normal	93.4	Outlier	82.1
	Dietary needs	Normal	129.5	Normal	143.2	Normal	6.9
	Dietary needs	Normal	82.9	Normal	163	Normal	34.3
	Dietary needs	Normal	97.8	Outlier	297.6	Outlier	70.7
	Dietary needs	Normal	50.9	Normal	112	Normal	19
	Dietary needs	Normal	99	Normal	104.7	Normal	37.7
	Dietary needs	Normal	136.1	Normal	131.6	Normal	18.9
	Dietary needs	Normal	138	Normal	115.2	Normal	2.7

Legumes Outliers?	Legumes
Normal	37.8
Normal	40.2
Normal	40.6
Normal	3.6
Normal	14.7
Normal	43.5
Normal	23.2
Normal	6.4
Normal	14.6
Normal	40.5
Normal	4.5
Normal	24.7
Normal	9.1
Normal	28.9
Normal	18.1
Normal	38.7
Normal	17.9
Normal	14
Normal	27.6
Normal	11.3
Normal	6.9
Normal	34.3
Normal	19
Normal	37.7
Normal	18.9
Normal	2.7
Normal	12.9
Normal	15.2
Normal	4.6
Normal	30.1
Normal	30.8

- Cleaning Country Dietary Needs Nuts variable:
  - Low outliers: Q1 (3.1) - IQR (9.5) x 1.5 = -11.15
  - High outliers: Q3 (12.6) + IQR (9.5) x 1.5 = 26.85

- Repeat steps from Fruit variable:

Table 38: Using IF statement to determine outliers for Nuts variable. Result of deselected outliers using excel filter on Nuts Outliers.

```
=IF(OR(O2 > 26.85, ISBLANK(O2)), "Outlier", "Normal")
```

J	K	L	M	N	O
Vegetables Outliers?	Vegetables	Legumes Outliers?	Legumes	Nuts Outliers?	Nuts
Normal	98.1	Outlier	103.8	Outlier	31.2
Outlier	308.8	Normal	37.8	Normal	11.6
Normal	129.1	Normal	40.2	Normal	12.3
Outlier		Outlier		Outlier	
Normal	123.3	Normal	40.6	Outlier	31.7
Normal	123.8	Normal	3.6	Normal	1.6
Normal	95.9	Normal	14.7	Normal	18.9
Normal	232.1	Normal	43.5	Normal	2.5
Normal	116.8	Normal	23.2	Normal	7.4
Normal	118.4	Normal	6.4	Normal	5.4
Normal	77.7	Normal	14.6	Outlier	32.7
Normal	186.8	Normal	40.5	Normal	6.5
Normal	113.2	Normal	4.5	Normal	3.4
Normal	99.1	Normal	24.7	Normal	13
Normal	27	Normal	9.1	Normal	21.8
Normal	225	Normal	28.9	Normal	0.7
Outlier	399.2	Normal	18.1	Normal	11.5
Normal	185.2	Normal	38.7	Normal	21.5
Normal	105.6	Normal	17.9	Normal	9.9
Outlier	443.6	Normal	14	Normal	8.8
Normal	87	Outlier	75.8	Normal	10.9
Normal	148.3	Normal	27.6	Normal	2.2
Normal	112.1	Normal	11.3	Normal	11.6
Normal	93.4	Outlier	82.1	Normal	0.2
Normal	143.2	Normal	6.9	Normal	8.7
Normal	163	Normal	34.3	Normal	4.2
Outlier	297.6	Outlier	70.7	Normal	6.6
Normal	112	Normal	19	Outlier	39.7
Normal	104.7	Normal	37.7	Normal	7.9
Normal	131.6	Normal	18.9	Normal	17.3
Normal	115.2	Normal	2.7	Normal	6.9

Nuts Outliers?	Nuts
Normal	11.6
Normal	12.3
Normal	1.6
Normal	18.9
Normal	2.5
Normal	7.4
Normal	5.4
Normal	6.5
Normal	3.4
Normal	13
Normal	21.8
Normal	0.7
Normal	11.5
Normal	21.5
Normal	9.9
Normal	8.8
Normal	10.9
Normal	2.2
Normal	11.6
Normal	0.2
Normal	8.7
Normal	4.2
Normal	6.6
Normal	7.9
Normal	17.3
Normal	6.9
Normal	17.9
Normal	9.4
Normal	9.6
Normal	14.9
Normal	26

- Cleaning Country Dietary Needs Whole grains variable:
  - Low outliers: Q1 (22) - IQR (35.2) x 1.5 = -30.8

- High outliers: Q3 (57.2) + IQR (35.2) x 1.5 = 110
- Repeat steps from Fruit variable:

Table 39: Using IF statement to determine outliers for Whole grains variable. Result of deselected outliers using excel filter on Whole grains Outliers.

```
=IF(OR(Q2 > 110, ISBLANK(Q2)), "Outlier", "Normal")
```

M	N	O	P	Q
Legumes	Nuts Outliers?	Nuts	Whole grains Outliers?	Whole grains
103.8	Outlier	31.2	Normal	41.5
37.8	Normal	11.6	Normal	57.6
40.2	Normal	12.3	Normal	31.2
	Outlier		Outlier	
40.6	Outlier	31.7	Normal	37.2
3.6	Normal	1.6	Normal	10.8
14.7	Normal	18.9	Normal	23.3
43.5	Normal	2.5	Normal	14.3
23.2	Normal	7.4	Normal	64.9
6.4	Normal	5.4	Normal	29.2
14.6	Outlier	32.7	Normal	46.9
40.5	Normal	6.5	Normal	33.1
4.5	Normal	3.4	Normal	66.1
24.7	Normal	13	Normal	33.6
9.1	Normal	21.8	Normal	59.4
28.9	Normal	0.7	Normal	35.3
18.1	Normal	11.5	Normal	25.8
38.7	Normal	21.5	Normal	16
17.9	Normal	9.9	Normal	36.7
14	Normal	8.8	Normal	45.7
75.8	Normal	10.9	Normal	35.7
27.6	Normal	2.2	Normal	27.6
11.3	Normal	11.6	Normal	29.4
82.1	Normal	0.2	Normal	6.1
6.9	Normal	8.7	Normal	89.9
34.3	Normal	4.2	Normal	0.1
70.7	Normal	6.6	Normal	45.4
19	Outlier	39.7	Normal	71.2
37.7	Normal	7.9	Normal	48.3
18.9	Normal	17.3	Normal	59.4
2.7	Normal	6.9	Normal	44.6

Whole grains Outliers?	Whole grains
Normal	41.5
Normal	57.6
Normal	31.2
Normal	37.2
Normal	10.8
Normal	23.3
Normal	14.3
Normal	64.9
Normal	29.2
Normal	46.9
Normal	33.1
Normal	66.1
Normal	33.6
Normal	59.4
Normal	35.3
Normal	25.8
Normal	16
Normal	36.7
Normal	45.7
Normal	35.7
Normal	27.6
Normal	29.4
Normal	6.1
Normal	89.9
Normal	0.1
Normal	45.4
Normal	71.2
Normal	48.3
Normal	59.4
Normal	44.6
Normal	40.2

- Cleaning Country Dietary Needs Fish variable:
  - Low outliers: Q1 (18.1) - IQR (26.3) x 1.5 = -21.35
  - High outliers: Q3 (44.4) + IQR (26.3) x 1.5 = 83.85
  - Repeat steps from Fruit variable:

Table 40: Using IF statement to determine outliers for Fish variable. Result of deselected outliers using excel filter on Fish Outliers.

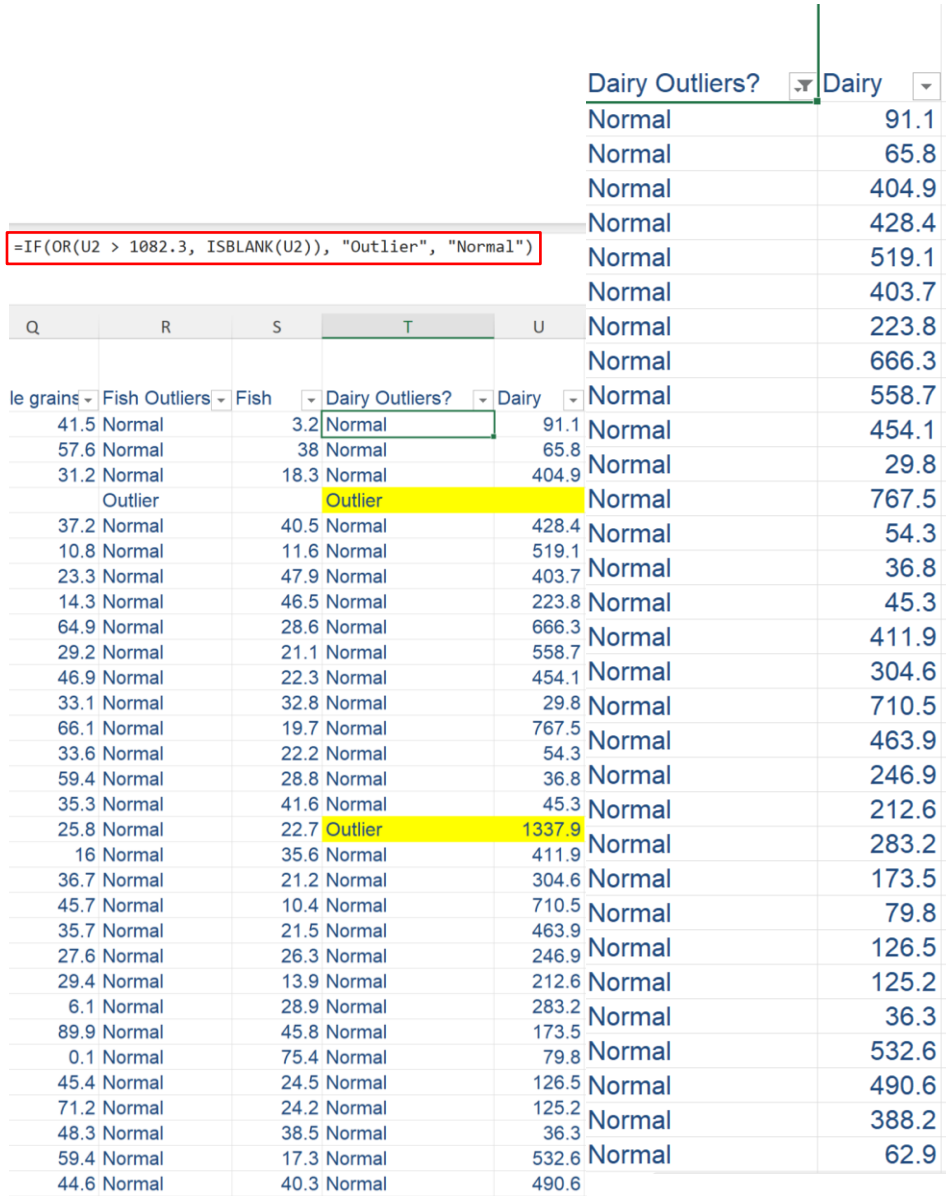
```
=IF(OR(S2 > 83.85, ISBLANK(S2)), "Outlier", "Normal")
```

P	Q	R	S	T
ole ns liers?	Whole grains	Fish Outliers	Fish	Dairy Outlier
mal	41.5	Normal	3.2	Normal
mal	57.6	Normal	38	Normal
mal	31.2	Normal	18.3	Normal
lier		Outlier		Outlier
mal	37.2	Normal	40.5	Normal
mal	10.8	Normal	11.6	Normal
mal	23.3	Normal	47.9	Normal
mal	14.3	Normal	46.5	Normal
mal	64.9	Normal	28.6	Normal
mal	29.2	Normal	21.1	Normal
mal	46.9	Normal	22.3	Normal
mal	33.1	Normal	32.8	Normal
mal	66.1	Normal	19.7	Normal
mal	33.6	Normal	22.2	Normal
mal	59.4	Normal	28.8	Normal
mal	35.3	Normal	41.6	Normal
mal	25.8	Normal	22.7	Outlier
mal	16	Normal	35.6	Normal
mal	36.7	Normal	21.2	Normal
mal	45.7	Normal	10.4	Normal
mal	35.7	Normal	21.5	Normal
mal	27.6	Normal	26.3	Normal
mal	29.4	Normal	13.9	Normal
mal	6.1	Normal	28.9	Normal
mal	89.9	Normal	45.8	Normal
mal	0.1	Normal	75.4	Normal
mal	45.4	Normal	24.5	Normal
mal	71.2	Normal	24.2	Normal
mal	48.3	Normal	38.5	Normal
mal	59.4	Normal	17.3	Normal
mal	44.6	Normal	40.3	Normal

Fish Outliers	Fish
Normal	3.2
Normal	38
Normal	18.3
Normal	40.5
Normal	11.6
Normal	47.9
Normal	46.5
Normal	28.6
Normal	21.1
Normal	22.3
Normal	32.8
Normal	19.7
Normal	22.2
Normal	28.8
Normal	41.6
Normal	22.7
Normal	35.6
Normal	21.2
Normal	10.4
Normal	21.5
Normal	26.3
Normal	13.9
Normal	28.9
Normal	45.8
Normal	75.4
Normal	24.5
Normal	24.2
Normal	38.5
Normal	17.3
Normal	40.3
Normal	16.6

- Cleaning Country Dietary Needs Dairy variable:
  - Low outliers: Q1 (80.3) - IQR (400.8) x 1.5 = -520.9
  - High outliers: Q3 (481.1) + IQR (400.8) x 1.5 = 1082.3
  - Repeat steps from Fruit variable:

Table 41: Using IF statement to determine outliers for Dairy variable. Result of deselected outliers using excel filter on Dairy Outliers.



```
=IF(OR(U2 > 1082.3, ISBLANK(U2)), "Outlier", "Normal")
```

Q	R	S	T	U
le grains	Fish Outliers	Fish	Dairy Outliers?	Dairy
41.5	Normal	3.2	Normal	91.1
57.6	Normal	38	Normal	65.8
31.2	Normal	18.3	Normal	404.9
	Outlier		Outlier	
37.2	Normal	40.5	Normal	428.4
10.8	Normal	11.6	Normal	519.1
23.3	Normal	47.9	Normal	403.7
14.3	Normal	46.5	Normal	223.8
64.9	Normal	28.6	Normal	666.3
29.2	Normal	21.1	Normal	558.7
46.9	Normal	22.3	Normal	454.1
33.1	Normal	32.8	Normal	29.8
66.1	Normal	19.7	Normal	767.5
33.6	Normal	22.2	Normal	54.3
59.4	Normal	28.8	Normal	36.8
35.3	Normal	41.6	Normal	45.3
25.8	Normal	22.7	Outlier	1337.9
16	Normal	35.6	Normal	411.9
36.7	Normal	21.2	Normal	304.6
45.7	Normal	10.4	Normal	710.5
35.7	Normal	21.5	Normal	463.9
27.6	Normal	26.3	Normal	246.9
29.4	Normal	13.9	Normal	212.6
6.1	Normal	28.9	Normal	283.2
89.9	Normal	45.8	Normal	173.5
0.1	Normal	75.4	Normal	79.8
45.4	Normal	24.5	Normal	126.5
71.2	Normal	24.2	Normal	125.2
48.3	Normal	38.5	Normal	36.3
59.4	Normal	17.3	Normal	532.6
44.6	Normal	40.3	Normal	490.6

Dairy Outliers?	Dairy
Normal	91.1
Normal	65.8
Normal	404.9
Normal	428.4
Normal	519.1
Normal	403.7
Normal	223.8
Normal	666.3
Normal	558.7
Normal	454.1
Normal	29.8
Normal	767.5
Normal	54.3
Normal	36.8
Normal	45.3
Normal	411.9
Normal	304.6
Normal	710.5
Normal	463.9
Normal	246.9
Normal	212.6
Normal	283.2
Normal	173.5
Normal	79.8
Normal	126.5
Normal	125.2
Normal	36.3
Normal	532.6
Normal	490.6
Normal	388.2
Normal	62.9

- Cleaning Country Dietary Needs Red meat variable:
  - Low outliers: Q1 (22.3) - IQR (44.2) x 1.5 = -44
  - High outliers: Q3 (66.5) + IQR (44.2) x 1.5 = 132.8
  - Repeat steps from Fruit variable:

Table 42: Using IF statement to determine outliers for Red meat variable. Result of deselected outliers using excel filter on Red meat Outliers.

```
=IF(OR(W2 > 132.8, ISBLANK(W2)), "Outlier", "Normal")
```

T	U	V	W
Dairy Outliers?	Dairy	Red meat Outliers?	Red mea
Normal	91.1	Normal	10.7
Normal	65.8	Normal	16.8
Normal	404.9	Normal	61.6
Outlier		Outlier	
Normal	428.4	Normal	29.6
Normal	519.1	Normal	86.7
Normal	403.7	Normal	62.4
Normal	223.8	Normal	36.3
Normal	666.3	Normal	61.6
Normal	558.7	Normal	106.2
Normal	454.1	Normal	99.3
Normal	29.8	Normal	11.8
Normal	767.5	Normal	57.4
Normal	54.3	Normal	16.5
Normal	36.8	Normal	2.7
Normal	45.3	Normal	8.1
Outlier	1337.9	Normal	73.3
Normal	411.9	Normal	42.3
Normal	304.6	Normal	53.5
Normal	710.5	Outlier	150.5
Normal	463.9	Normal	61.9
Normal	246.9	Normal	98.7
Normal	212.6	Normal	99.5
Normal	283.2	Normal	75.2
Normal	173.5	Normal	64.8
Normal	79.8	Normal	90.8
Normal	126.5	Normal	17.9
Normal	125.2	Normal	25
Normal	36.3	Normal	80
Normal	532.6	Normal	46.4
Normal	490.6	Normal	49.3

Red meat Outliers?	Red mea
Normal	10.7
Normal	16.8
Normal	61.6
Normal	29.6
Normal	86.7
Normal	62.4
Normal	36.3
Normal	61.6
Normal	106.2
Normal	99.3
Normal	11.8
Normal	57.4
Normal	16.5
Normal	2.7
Normal	8.1
Normal	73.3
Normal	42.3
Normal	53.5
Normal	61.9
Normal	98.7
Normal	99.5
Normal	75.2
Normal	64.8
Normal	90.8
Normal	17.9
Normal	25
Normal	80
Normal	46.4
Normal	49.3
Normal	18.6
Normal	105

- To address missing values, outliers will be excluded, and a filtered mean will be used to fill in missing values for the Country Dietary Needs datasheet.
    - Addressing missing values for the Fruit Variable by filling in missing values with the filtered mean, 91.73895349:

Table 43: Deselecting outlier on the Fruit Outliers tab to display values without outliers.

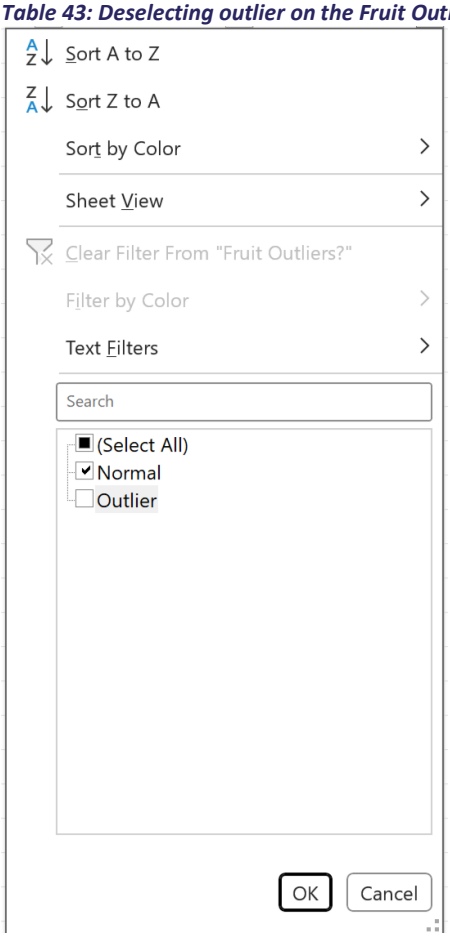


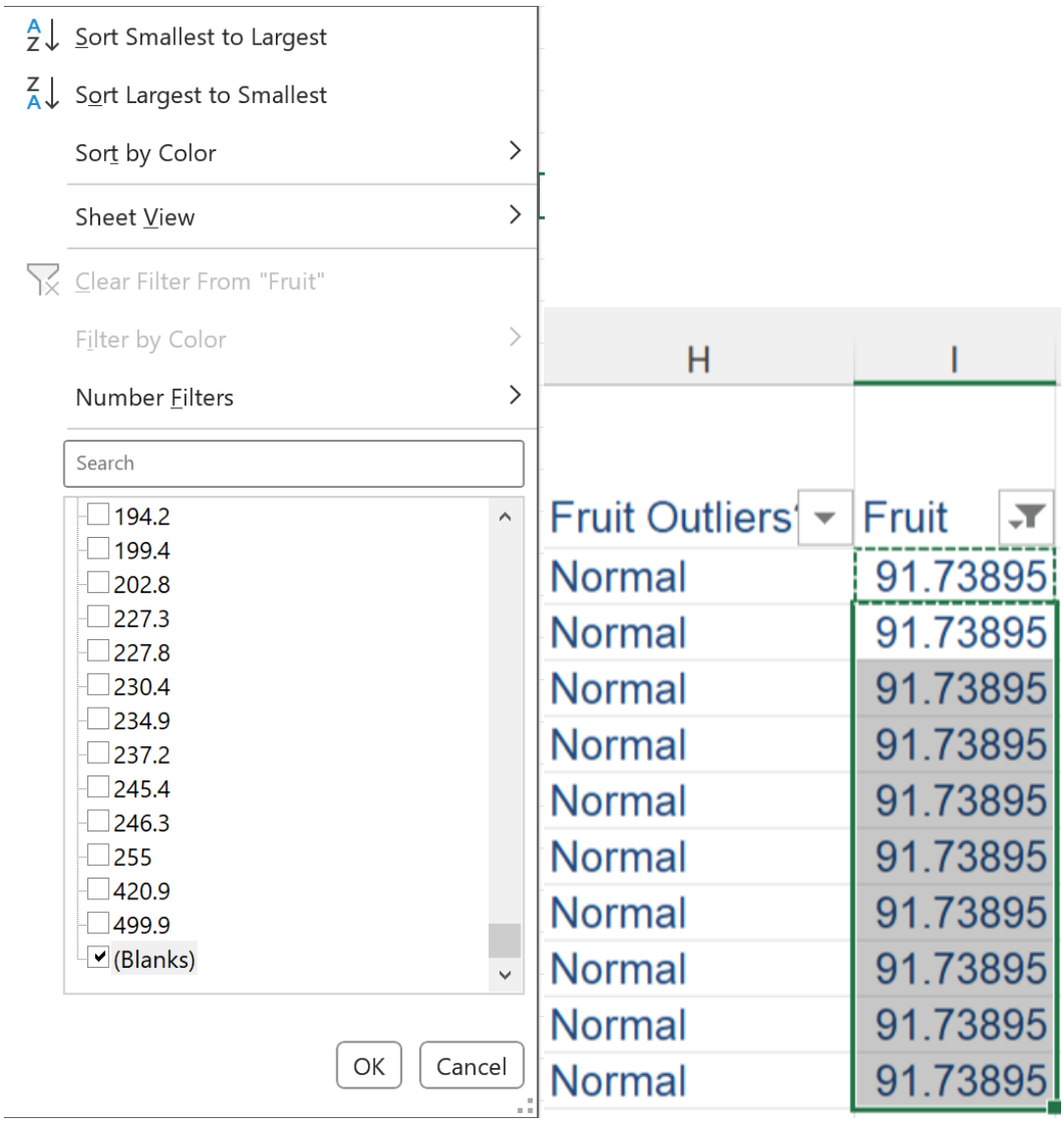
Table 44: Results of filtered outliers to calculate filtered mean to apply to missing values. The filtered mean is the average near the bottom right of the figure.

	H	I	J	K	L	M	N
	Fruit Outliers	Fruit	Vegetables Outliers?	Vegetables	Legumes Outliers?	Legume	Nuts Outliers?
	Normal	65.7	Normal	98.1	Outlier	103.8	Outlier
	Normal	119.7	Outlier	308.8	Normal	37.8	Normal
	Normal	138.9	Normal	129.1	Normal	40.2	Normal
	Normal	107.6	Normal	123.3	Normal	40.6	Outlier
	Normal	94	Normal	123.8	Normal	3.6	Normal
	Normal	91.2	Normal	95.9	Normal	14.7	Normal
	Normal	95.5	Normal	232.1	Normal	43.5	Normal
	Normal	131.9	Normal	116.8	Normal	23.2	Normal
	Normal	105.7	Normal	118.4	Normal	6.4	Normal
	Normal	104.8	Normal	77.7	Normal	14.6	Outlier
	Normal	73.8	Normal	186.8	Normal	40.5	Normal
	Normal	103.9	Normal	113.2	Normal	4.5	Normal
	Normal	74.8	Normal	99.1	Normal	24.7	Normal
	Normal	34.8	Normal	27	Normal	9.1	Normal
	Normal	54.7	Normal	225	Normal	28.9	Normal
	Normal	95.9	Outlier	399.2	Normal	18.1	Normal
	Normal	119.8	Normal	185.2	Normal	38.7	Normal
	Normal	83.7	Normal	105.6	Normal	17.9	Normal
	Normal	85.6	Normal	148.3	Normal	27.6	Normal
	Normal	86.6	Normal	112.1	Normal	11.3	Normal

ary needs | Region dietary needs | Country interventio ...

Average: 91.73895349 Count: 173 Sum: 15779.1

Table 45: Deselecting all values except missing values to determine the missing data that will use the filtered mean. Filled in the missing values with the filtered mean.



- Addressing missing values for the Vegetables Variable by filling in missing values with the filtered mean, 132.8052941:

Table 46: Deselecting outlier on the Vegetables Outliers tab to display values without outliers alongside results of filtered outliers to calculate filtered mean to apply to missing values. The filtered mean is the average near the bottom right of the figure.

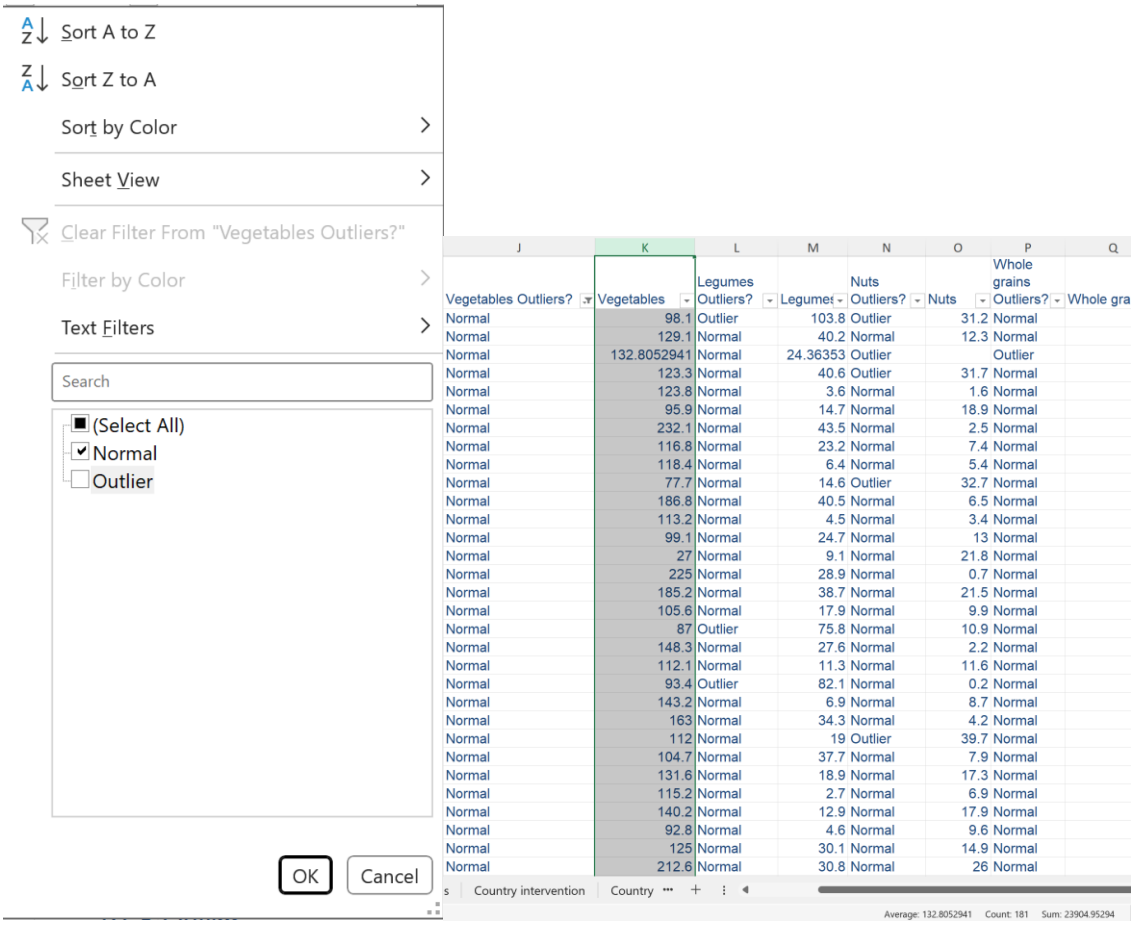
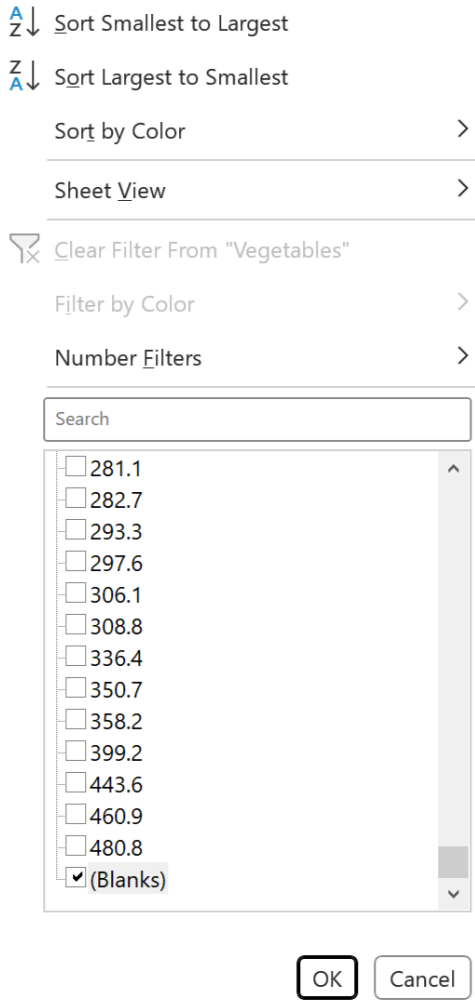


Table 47: Deselecting all values except missing values to determine the missing data that will use the filtered mean. Filled in the missing values with the filtered mean.



Vegetables Outliers?	Vegetables
Normal	132.8052941
Normal	132.8052941
Normal	132.8052941
Normal	132.8052941
Normal	132.8052941
Normal	132.8052941
Normal	132.8052941
Normal	132.8052941
Normal	132.8052941
Normal	132.8052941

- Addressing missing values for the Legumes Variable by filling in missing values with the filtered mean, 24.36352941:

Table 48: Deselecting outlier on the Legumes Outliers tab to display values without outliers alongside results of filtered outliers to calculate filtered mean to apply to missing values. The filtered mean is the average near the bottom right of the figure.

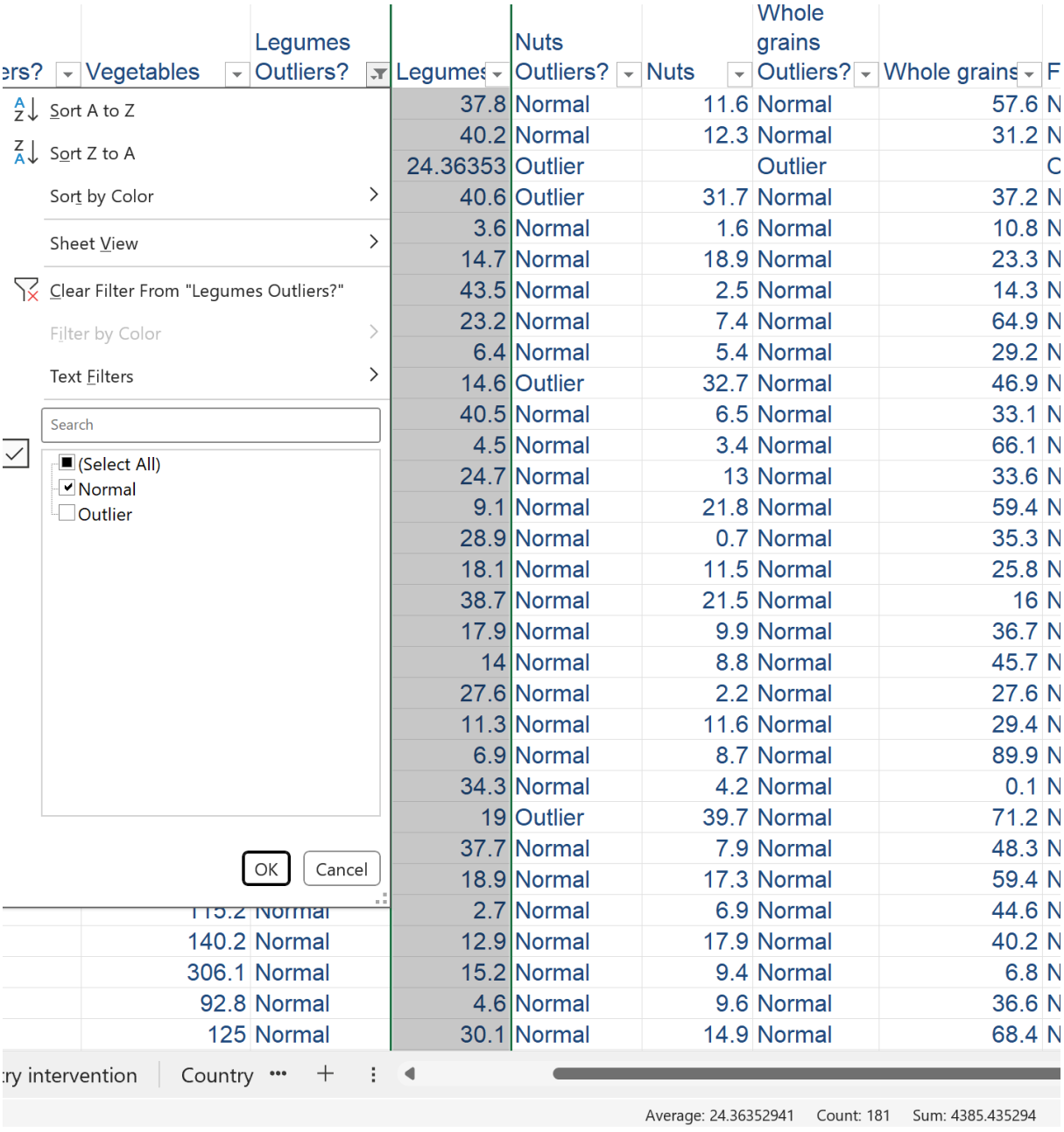
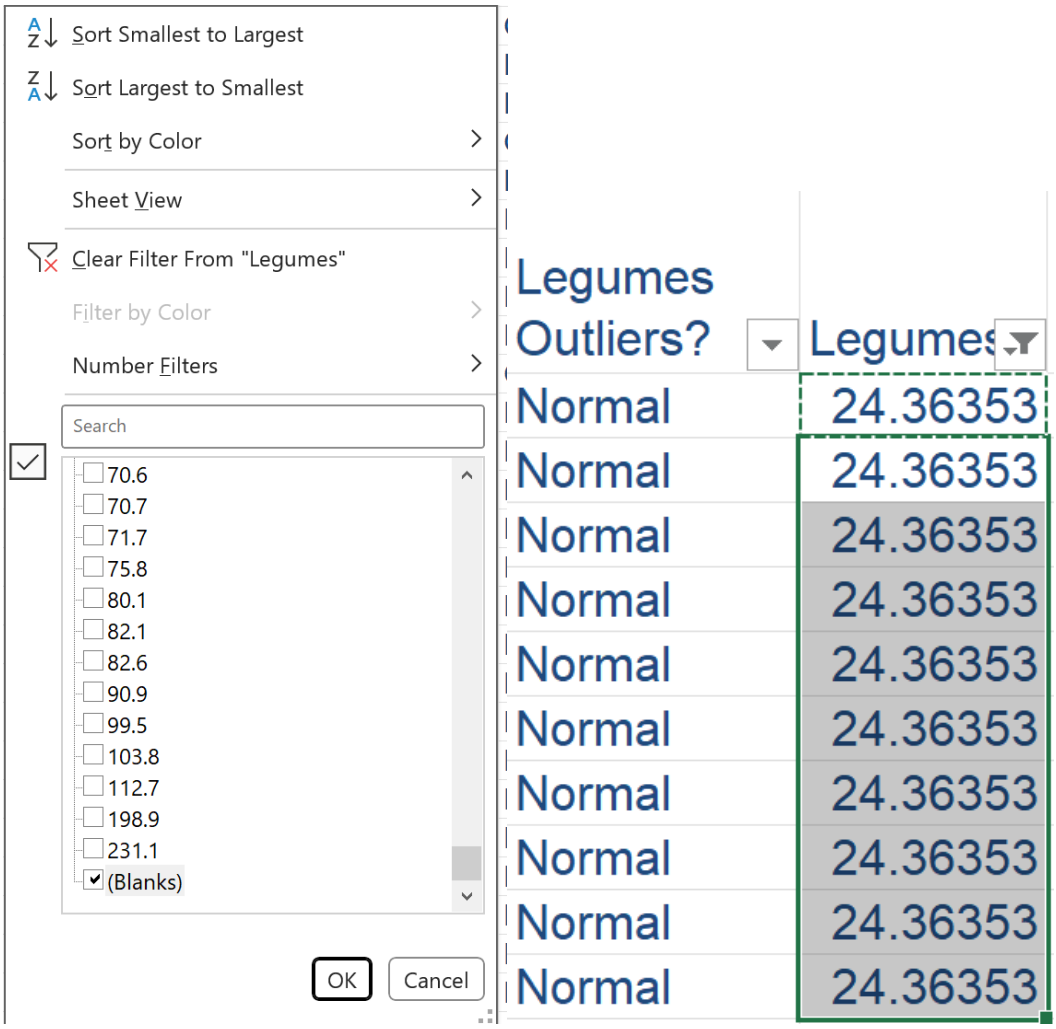


Table 49: Deselecting all values except missing values to determine the missing data that will use the filtered mean. Filled in the missing values with the filtered mean.



- Addressing missing values for the Nuts Variable by filling in missing values with the filtered mean, 7.94127907:

Table 50: Deselecting outlier on the Nuts Outliers tab to display values without outliers alongside results of filtered outliers to calculate filtered mean to apply to missing values. The filtered mean is the average near the bottom right of the figure.

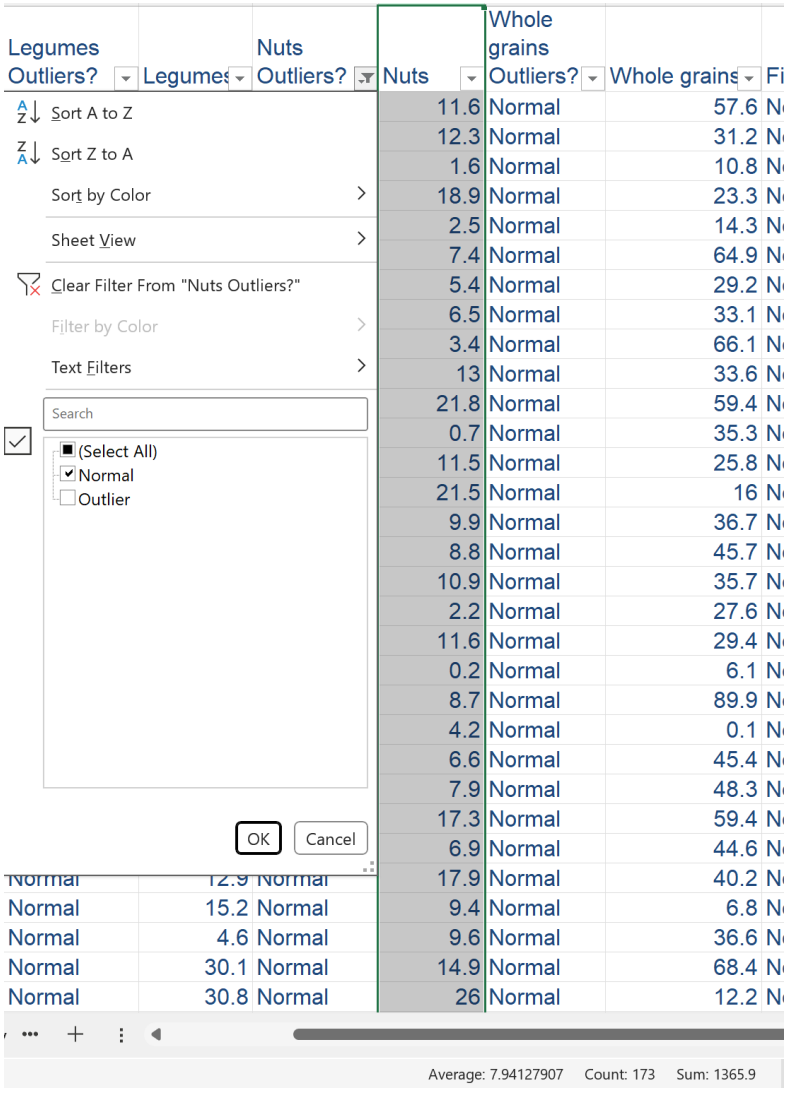
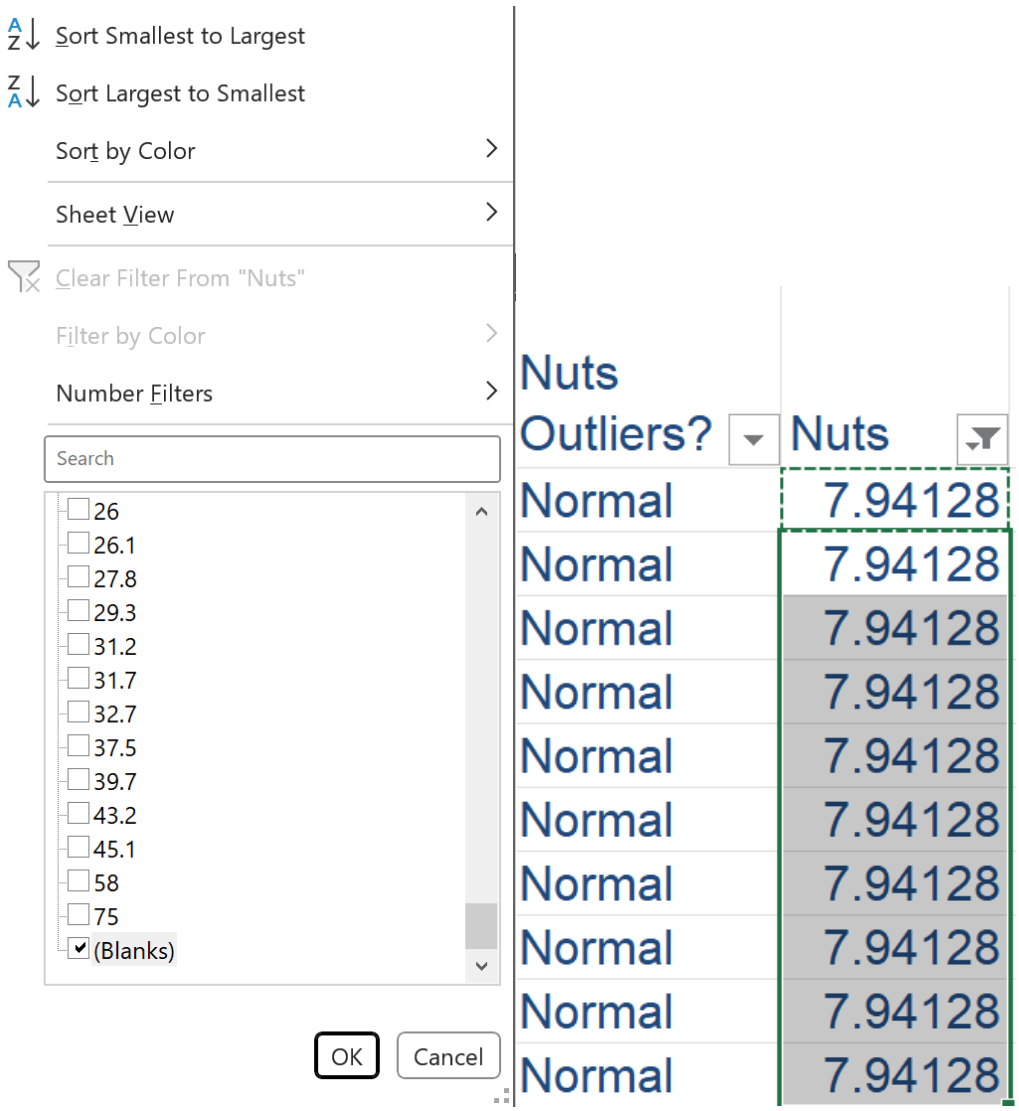


Table 51: Deselecting all values except missing values to determine the missing data that will use the filtered mean. Filled in the missing values with the filtered mean.



- Addressing missing values for the Whole grains Variable by filling in missing values with the filtered mean, 36.87861272:

Table 52: Deselecting outlier on the Whole grains Outliers tab to display values without outliers alongside results of filtered outliers to calculate filtered mean to apply to missing values. The filtered mean is the average near the bottom right of the figure.

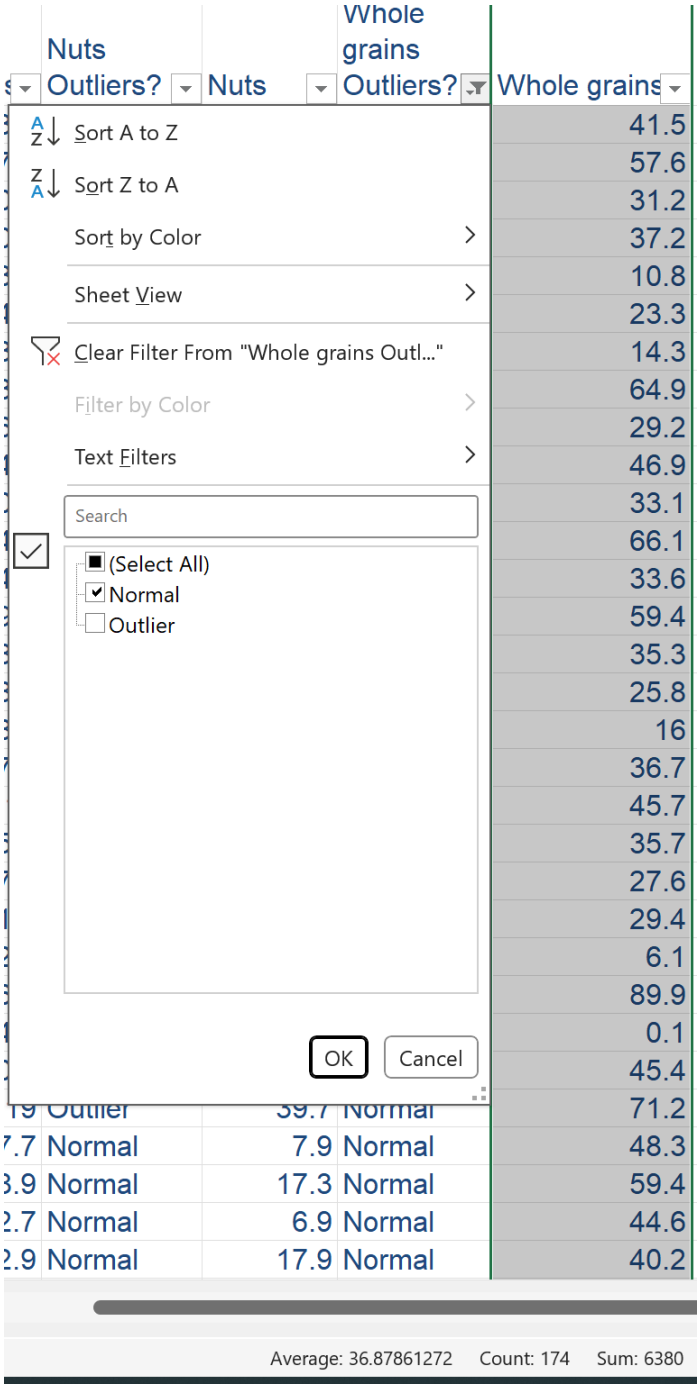
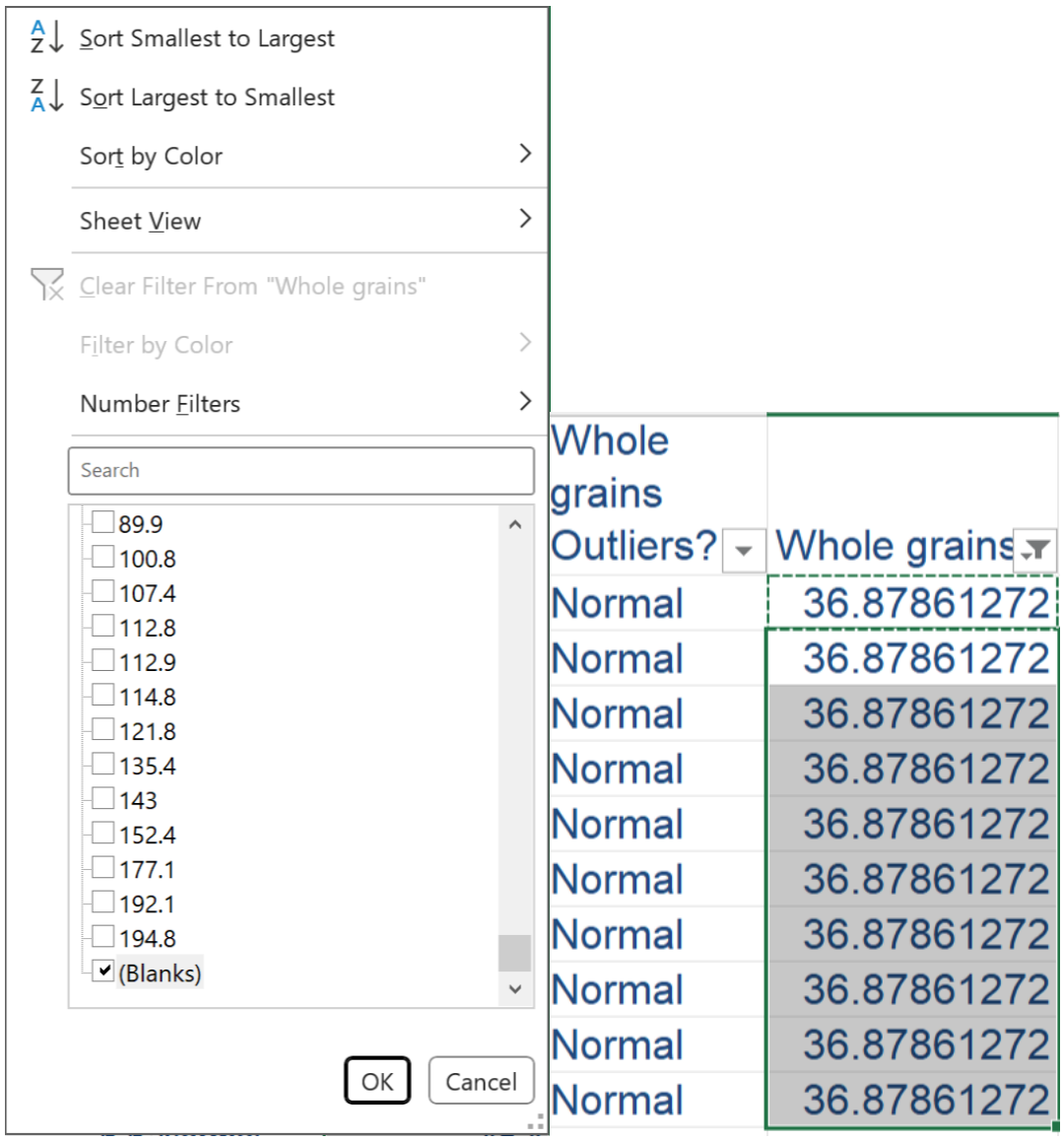


Table 53: Deselecting all values except missing values to determine the missing data that will use the filtered mean. Filled in the missing values with the filtered mean.



- Addressing missing values for the Fish Variable by filling in missing values with the filtered mean, 31.5752809:

Table 54: Deselecting outlier on the Fish Outliers tab to display values without outliers alongside results of filtered outliers to calculate filtered mean to apply to missing values. The filtered mean is the average near the bottom right of the figure.

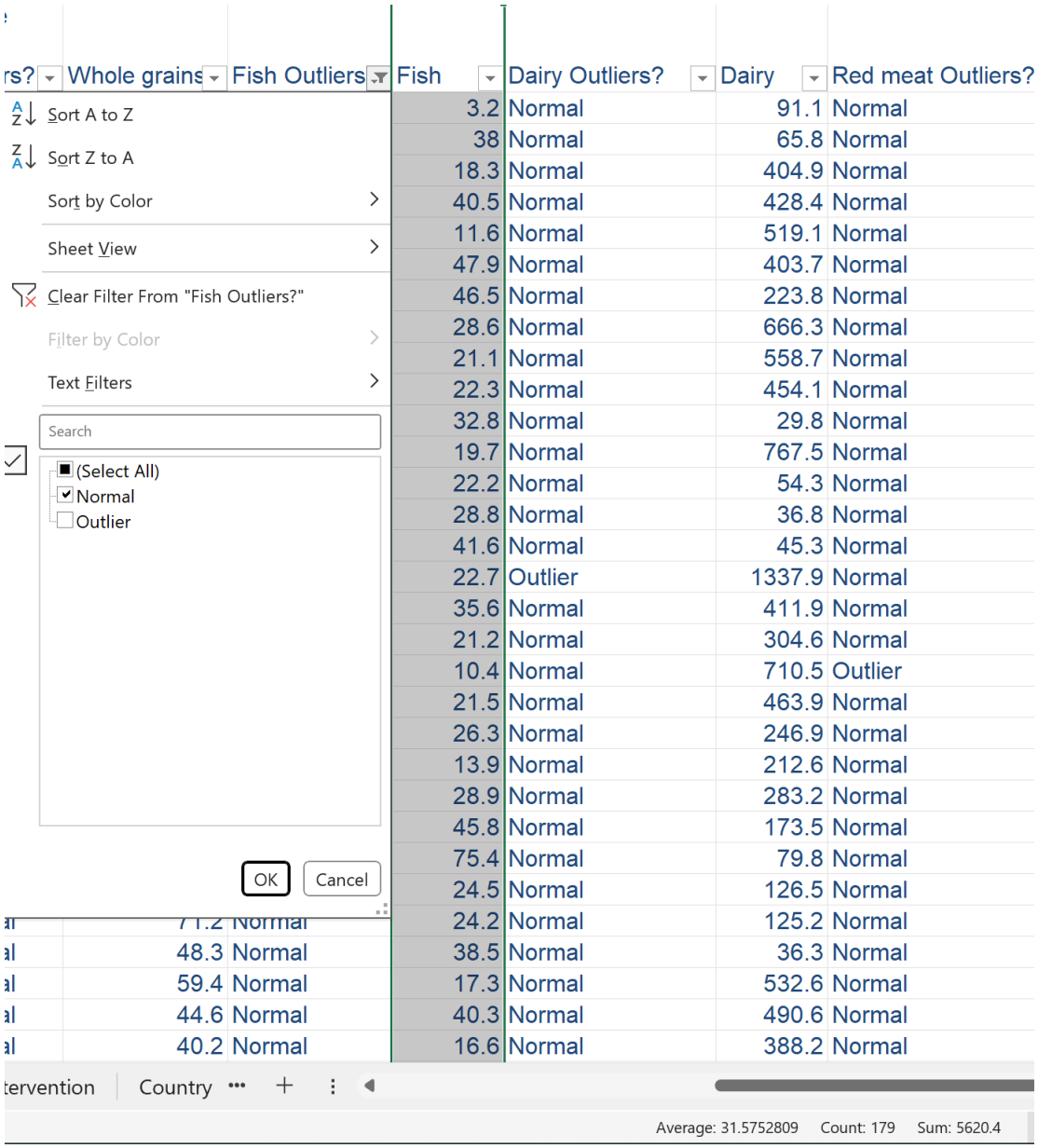
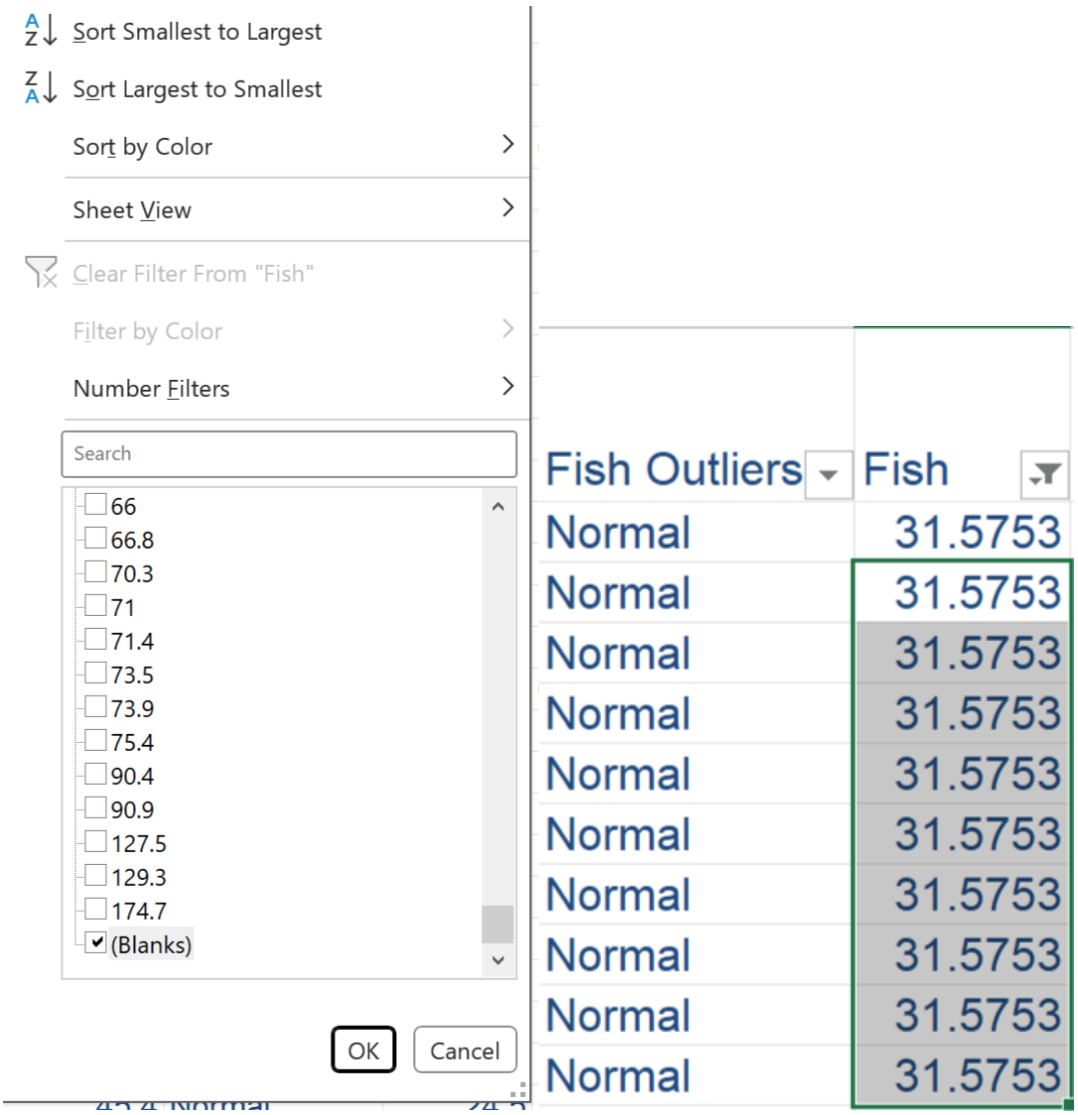


Table 55: Deselecting all values except missing values to determine the missing data that will use the filtered mean. Filled in the missing values with the filtered mean.



- Addressing missing values for the Dairy Variable by filling in missing values with the filtered mean, 292.3572222:

Table 56: Deselecting outlier on the Dairy Outliers tab to display values without outliers alongside results of filtered outliers to calculate filtered mean to apply to missing values. The filtered mean is the average near the bottom right of the figure.

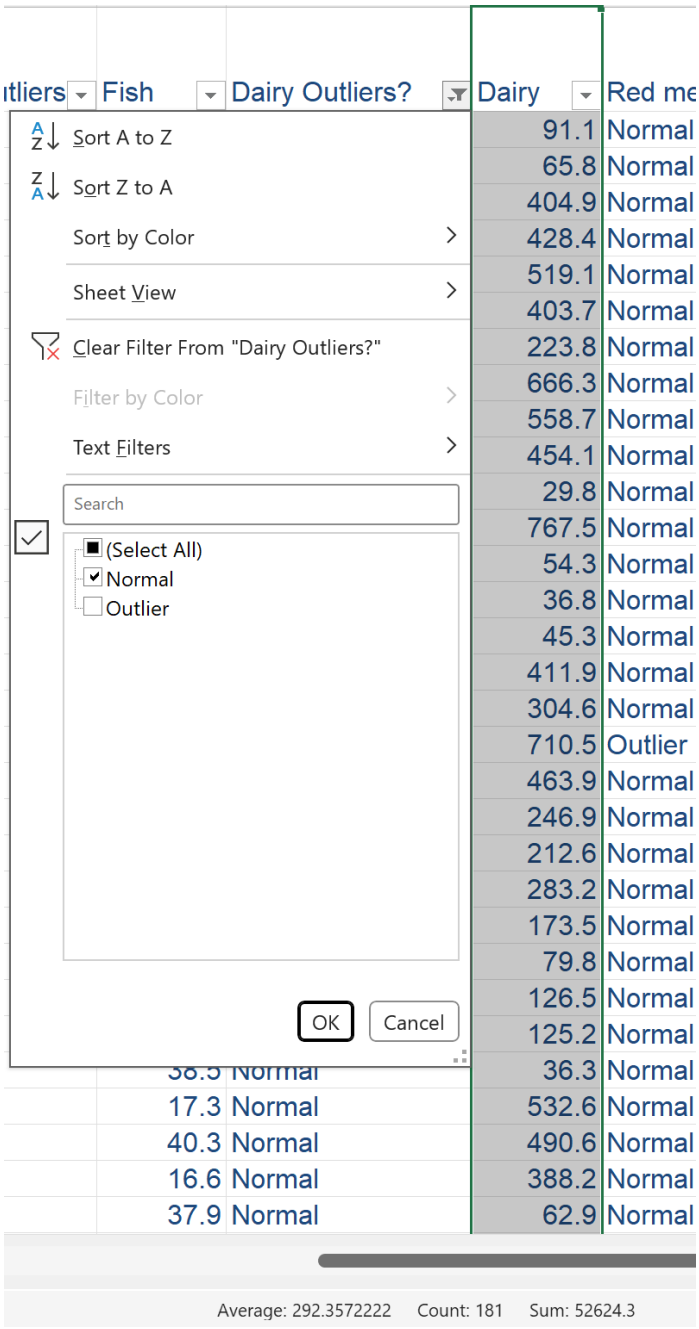
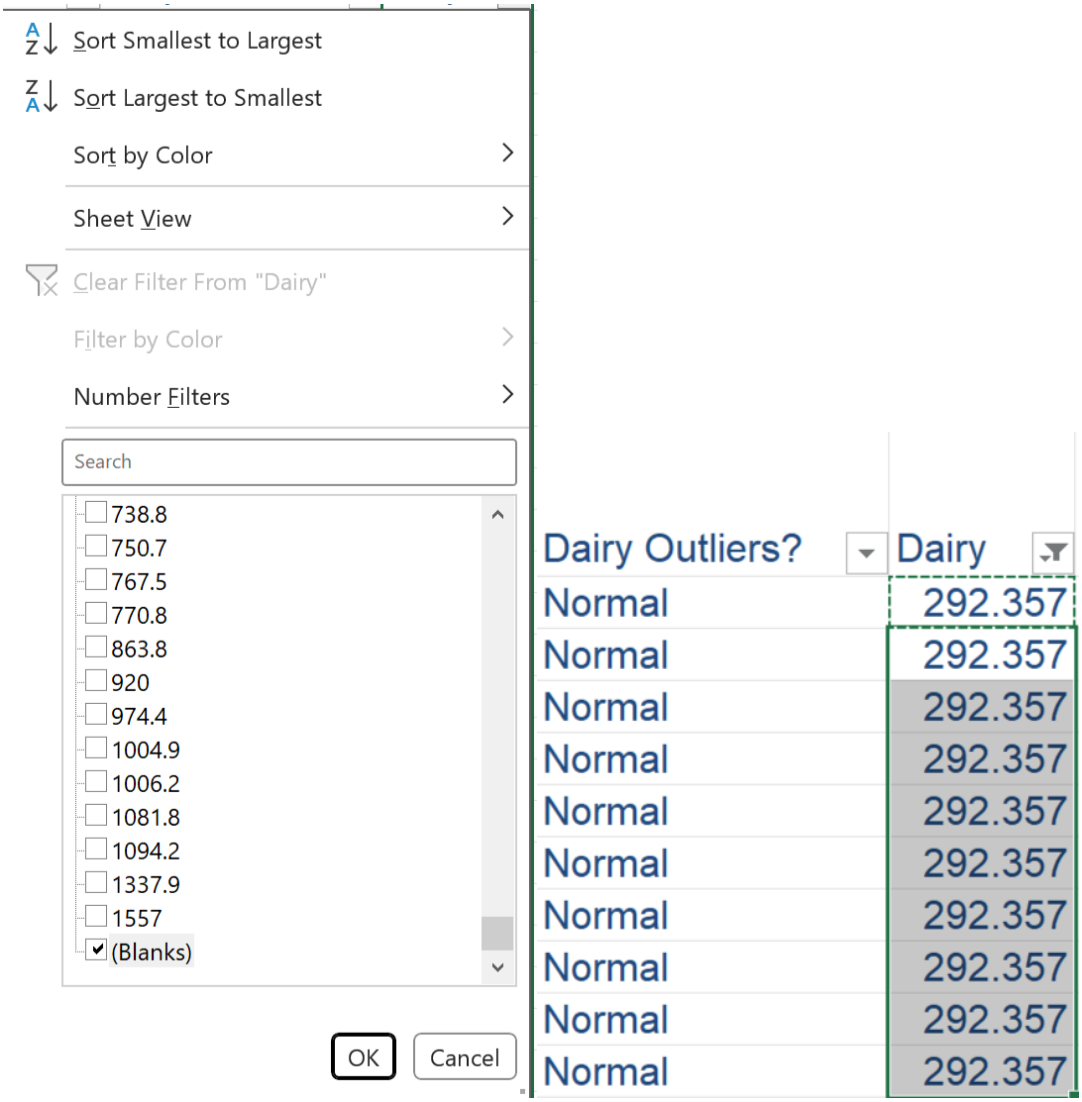


Table 57: Deselecting all values except missing values to determine the missing data that will use the filtered mean. Filled in the missing values with the filtered mean.



- Addressing missing values for the Red meat Variable by filling in missing values with the filtered mean, 45.03828571:

Table 58: Deselecting outlier on the Red meat Outliers tab to display values without outliers alongside results of filtered outliers to calculate filtered mean to apply to missing values. The filtered mean is the average near the bottom right of the figure.

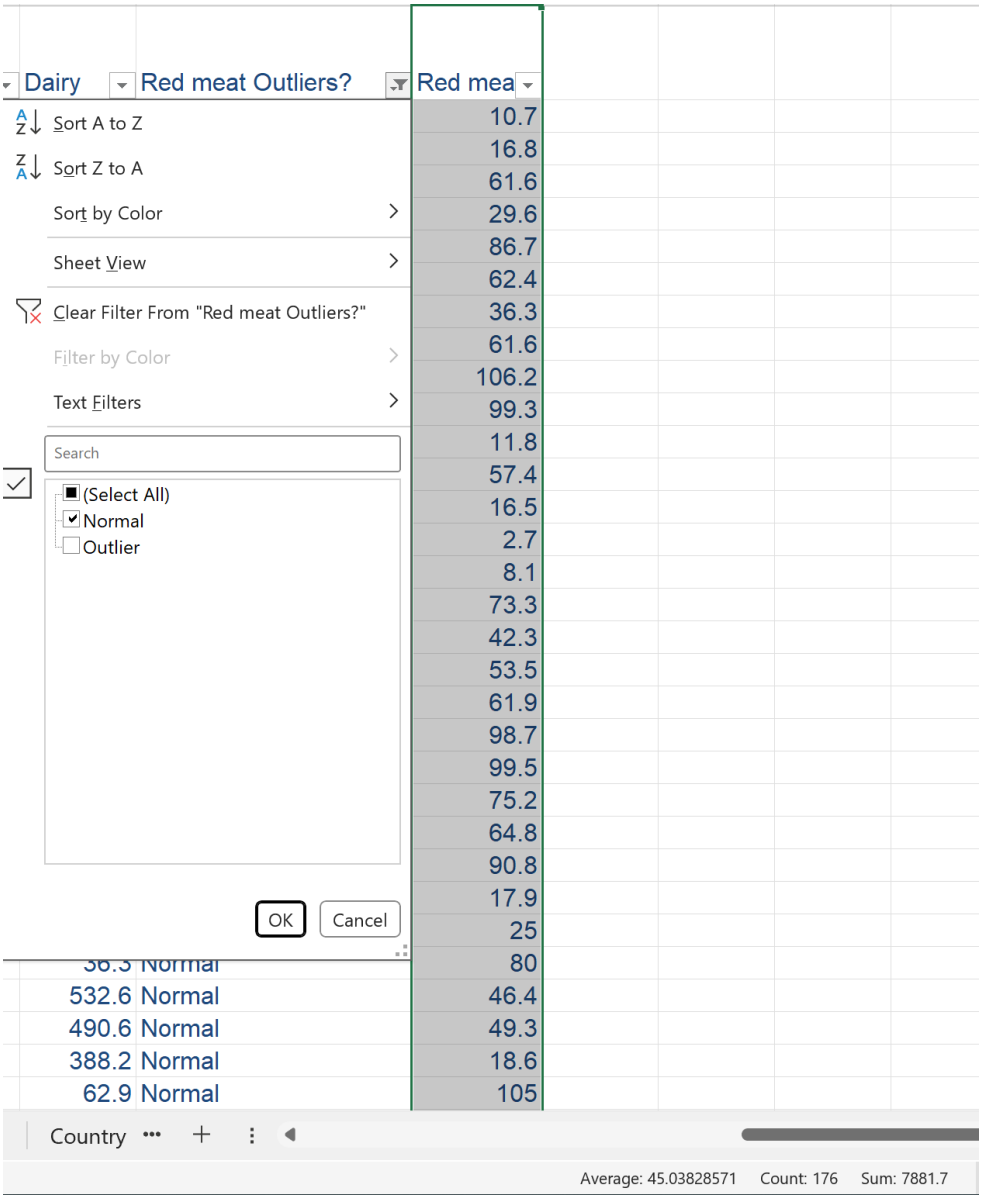
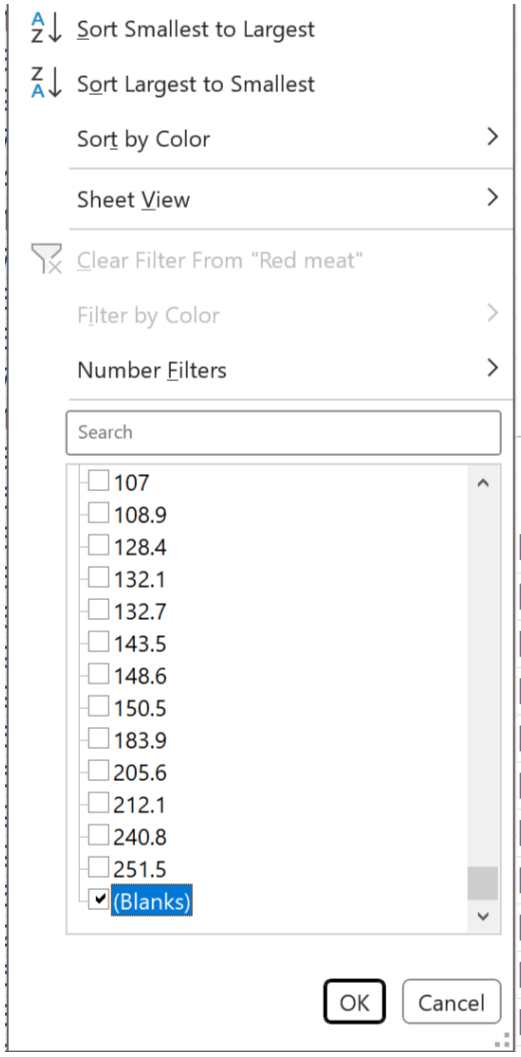


Table 59: Deselecting all values except missing values to determine the missing data that will use the filtered mean. Filled in the missing values with the filtered mean.



Red meat Outliers?	Red mea
Normal	45.03829
Normal	45.03829
Normal	45.03829
Normal	45.03829
Normal	45.03829
Normal	45.03829
Normal	45.03829
Normal	45.03829
Normal	45.03829
Normal	45.03829

Modeling

Modeling Techniques

Descriptive Models:

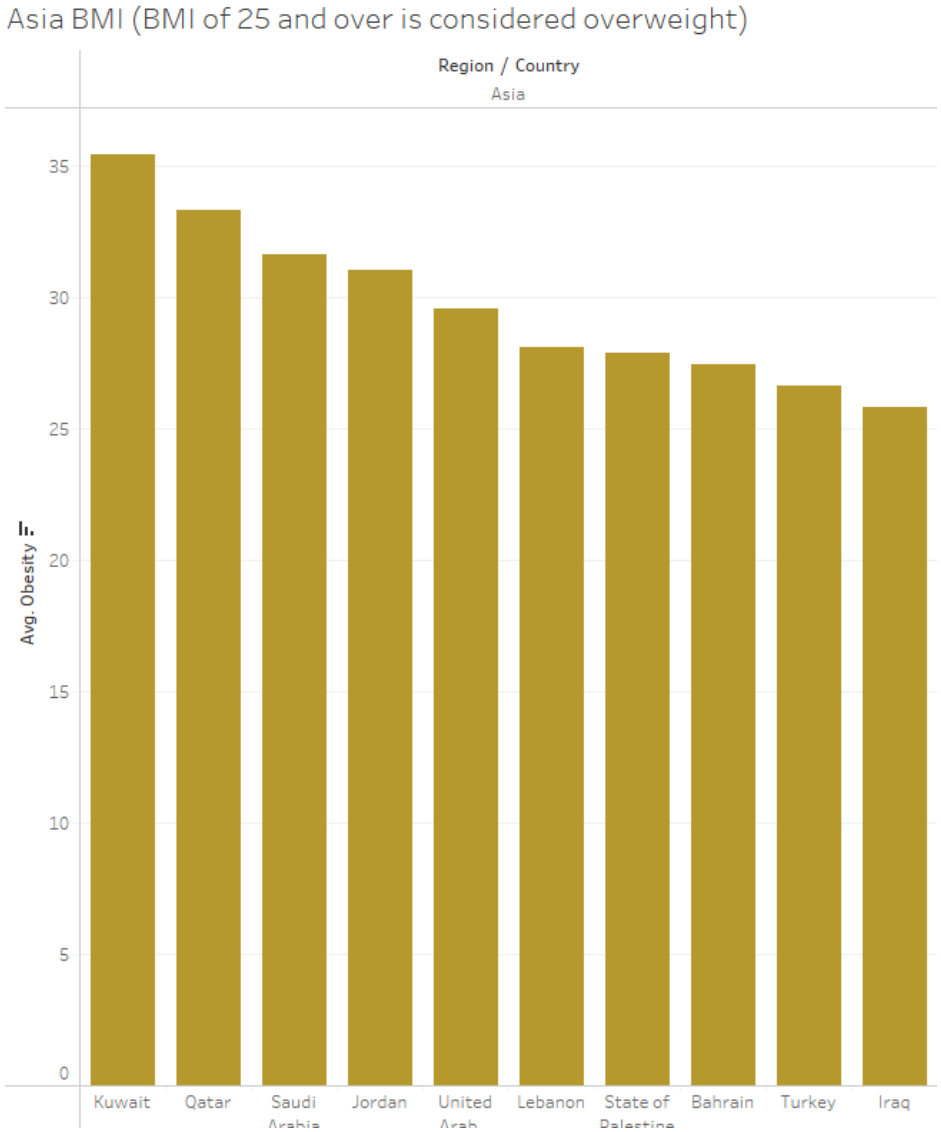


Figure 15: This bar graph shows Top 10 Asian countries with the highest obesity rates (calculated by average body mass index).

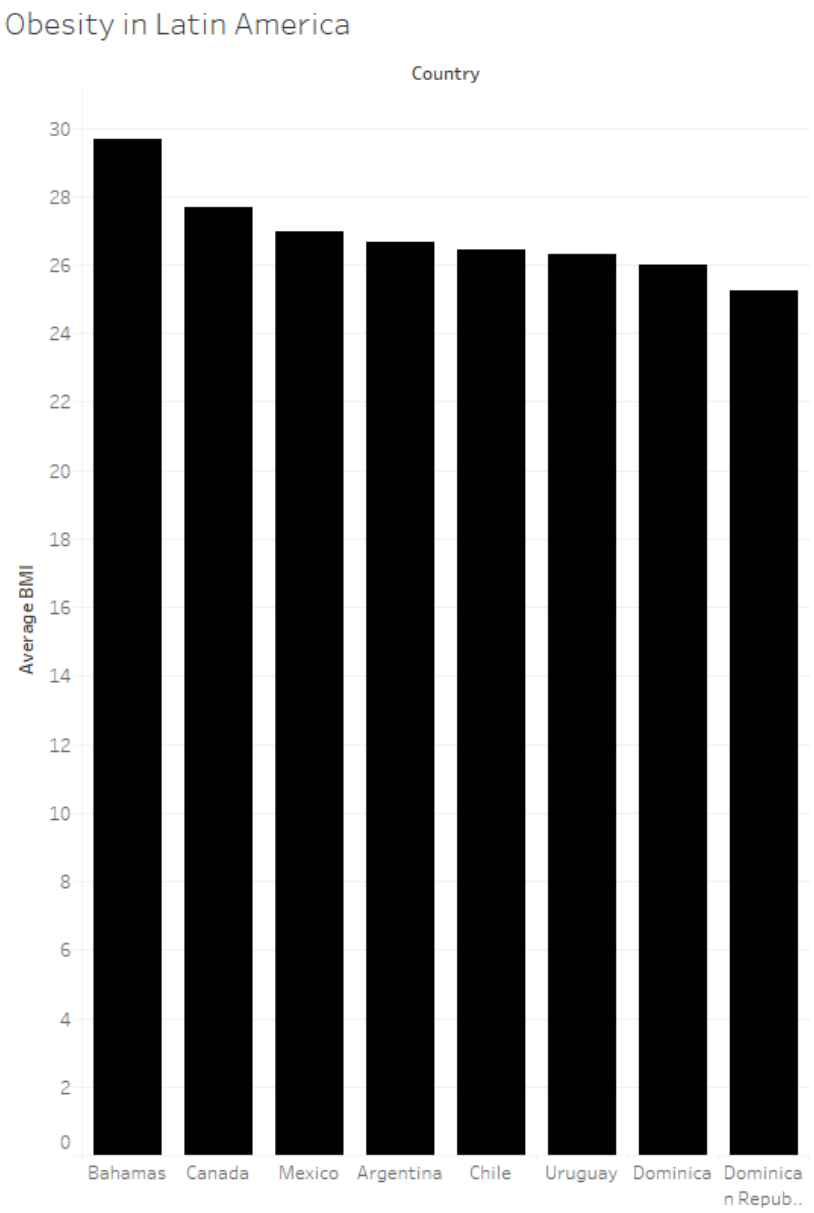


Figure 16: This bar graph shows Top 10 Latin American countries with the highest obesity rates (calculated by average body mass index).

- According to the World Health Organization (WHO) a BMI of 25 and higher is considered overweight.

In this data set we focused on two regions that showed higher obesity rates: North and Latin America and Asia. From These two regions, we selected the countries that best align with Dollar General's goals in improving accessible healthy foods.

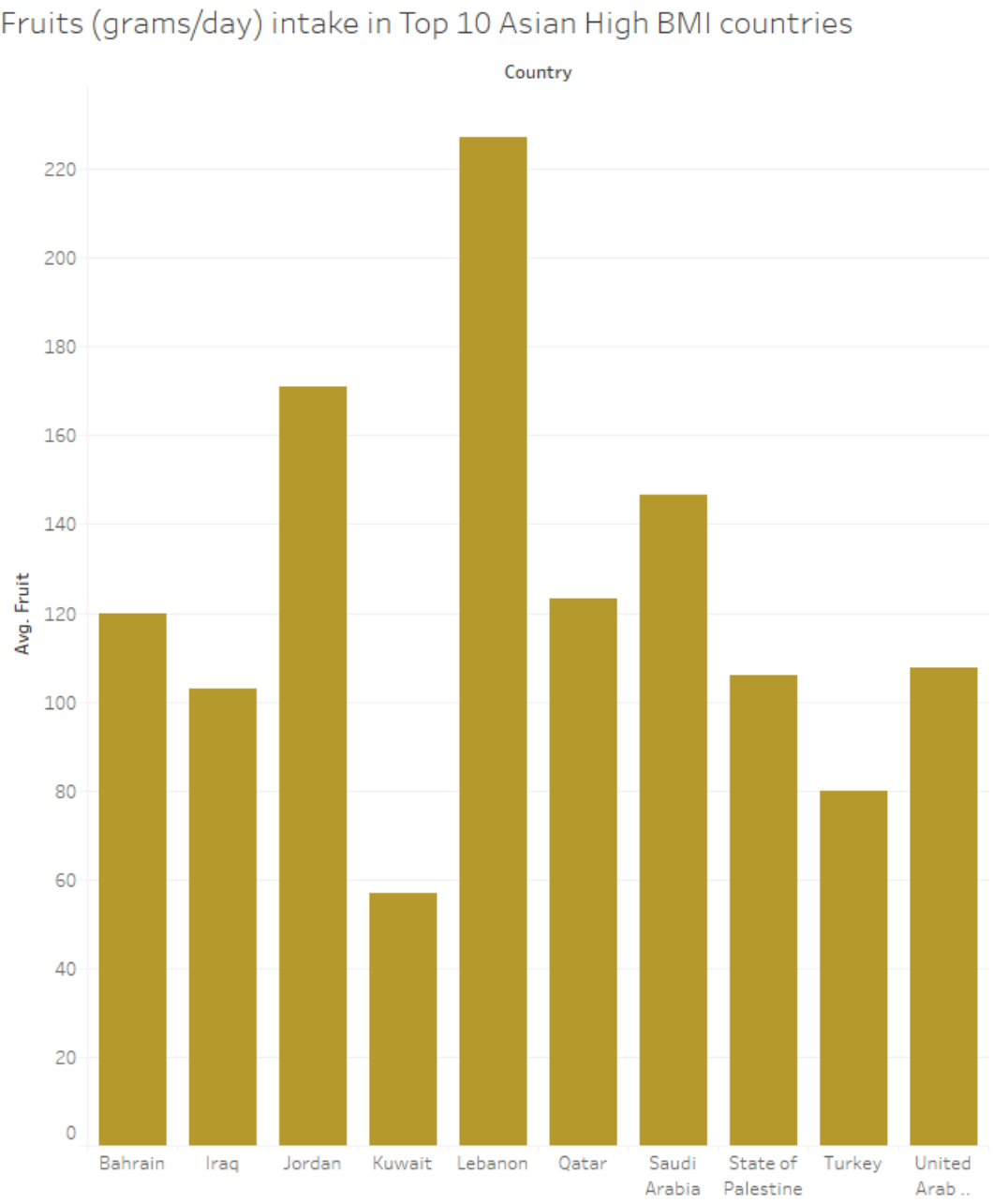


Figure 17: **The bar graph shows fruit intake (grams/day) for the top 10 Asian countries with high obesity.**

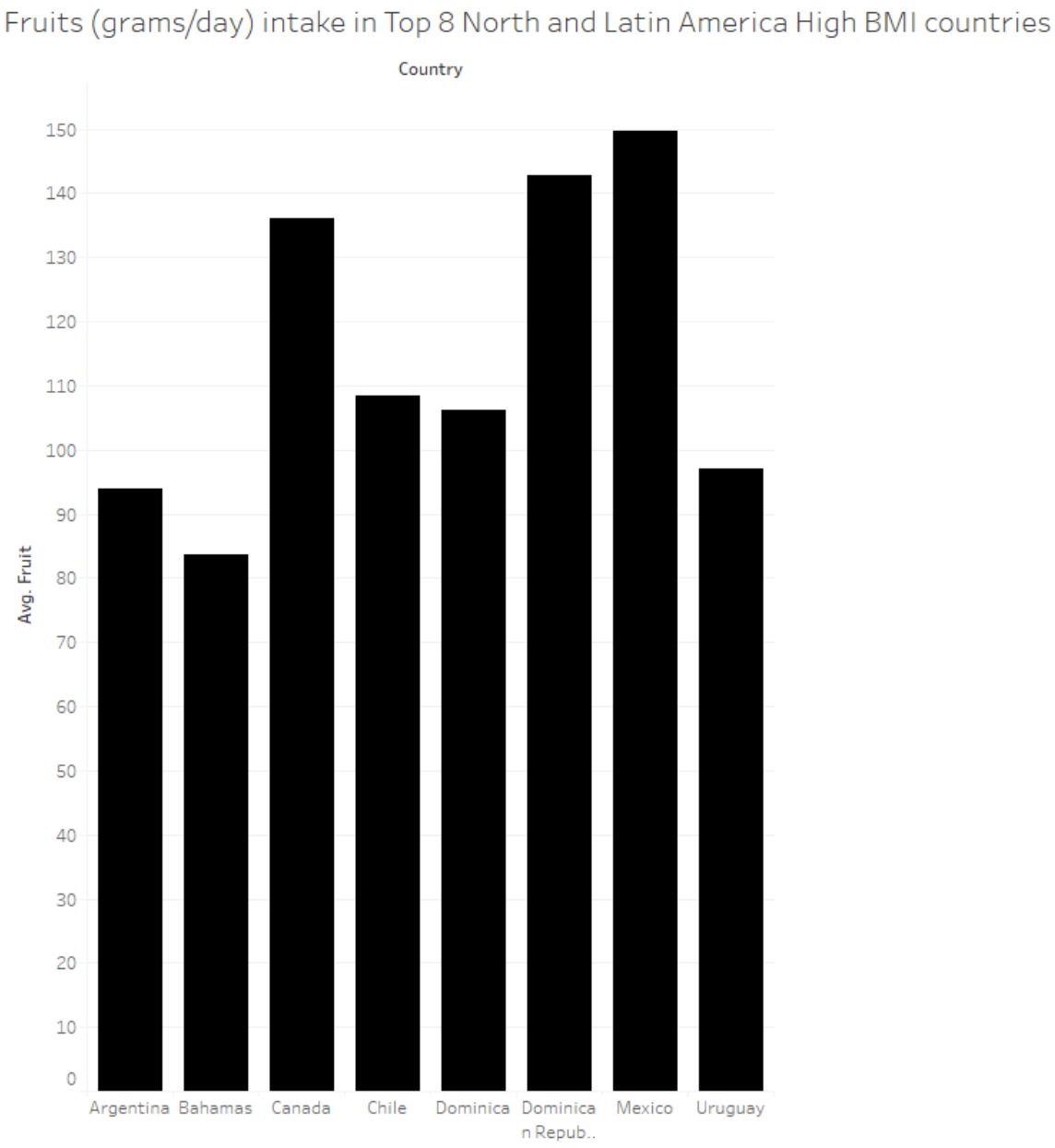


Figure 18: The bar graph shows fruit intake (grams/day) for the top 8 Latin American countries with high obesity.

We noticed and understood that fruit and vegetable underconsumption are linked to obesity, upon our findings, no country was meeting recommended guidelines. According to the World Health Organization (WHO) the recommended daily average fruit intake is 400 grams. Most of the countries fall between 150 more or less grams. Fruits impact reduces obesity by its low energy density, high fiber, and satiety effect.

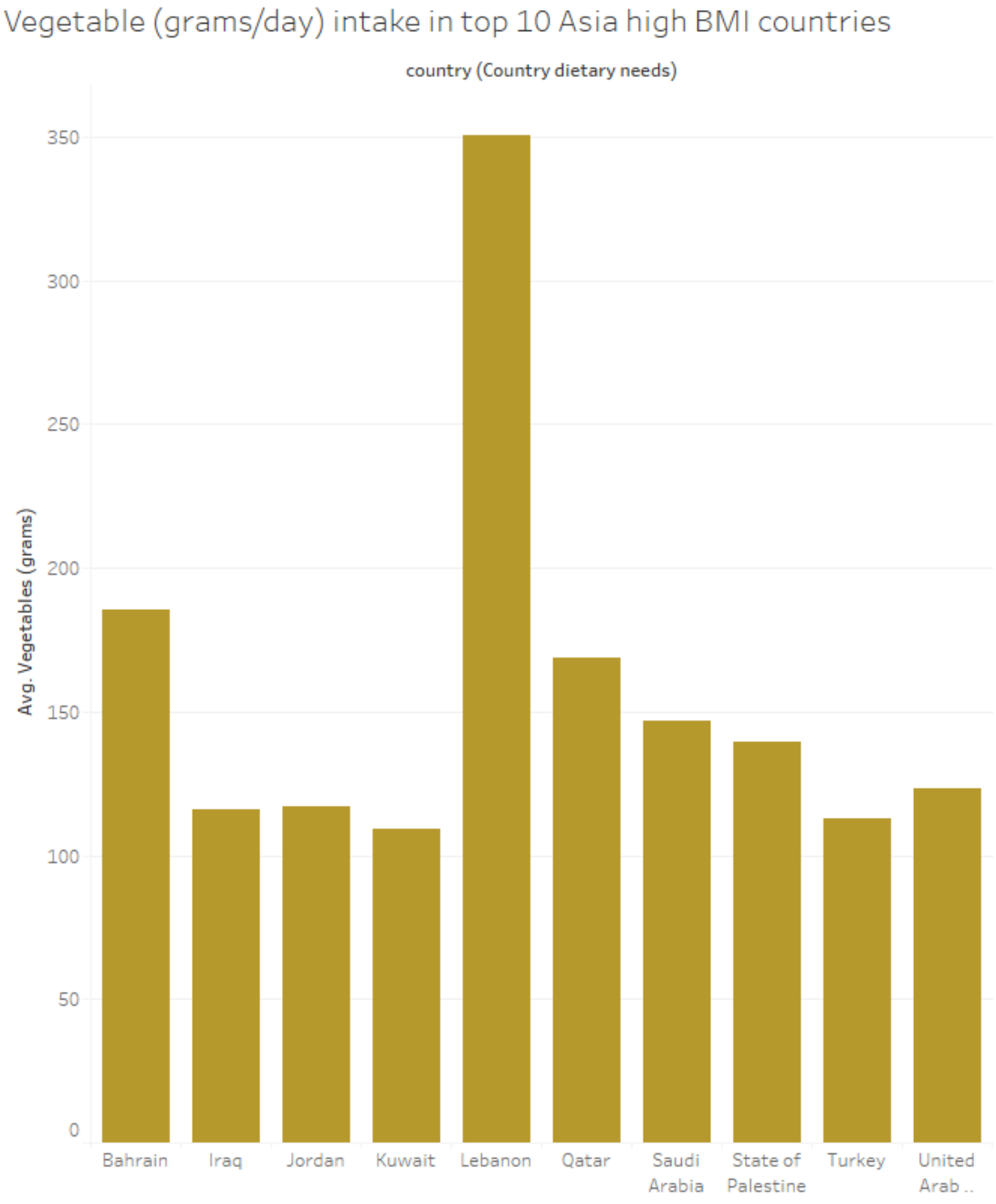


Figure 19: The bar graph shows vegetable intake (grams/day) for the top 10 Asian countries with high obesity.

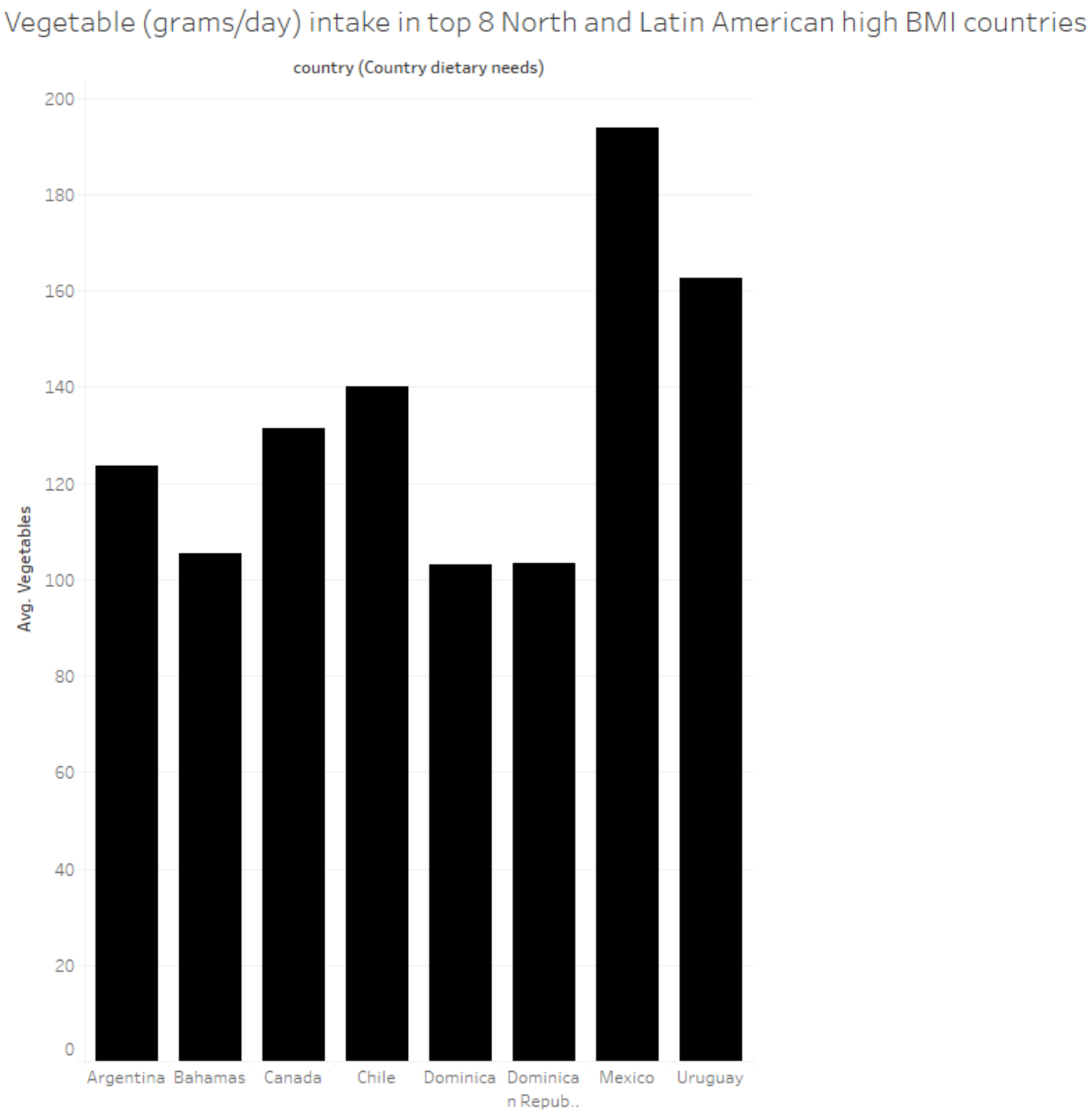


Figure 20: The bar graph shows vegetable intake (grams/day) for the top 8 Latin American countries with high obesity.

Underconsumption of Vegetables are linked to high obesity rates. In the findings by Department of Community Nutrition in Shiraz University of Medical Sciences found that obesity decreases obesity by 21% if two servings of vegetables are consumed daily. For fruits obesity decreases by 26% with two servings daily (Mehran Nouri, 2023, p. 5). The recommended daily intake of vegetables is 150 grams per day according to WHO.

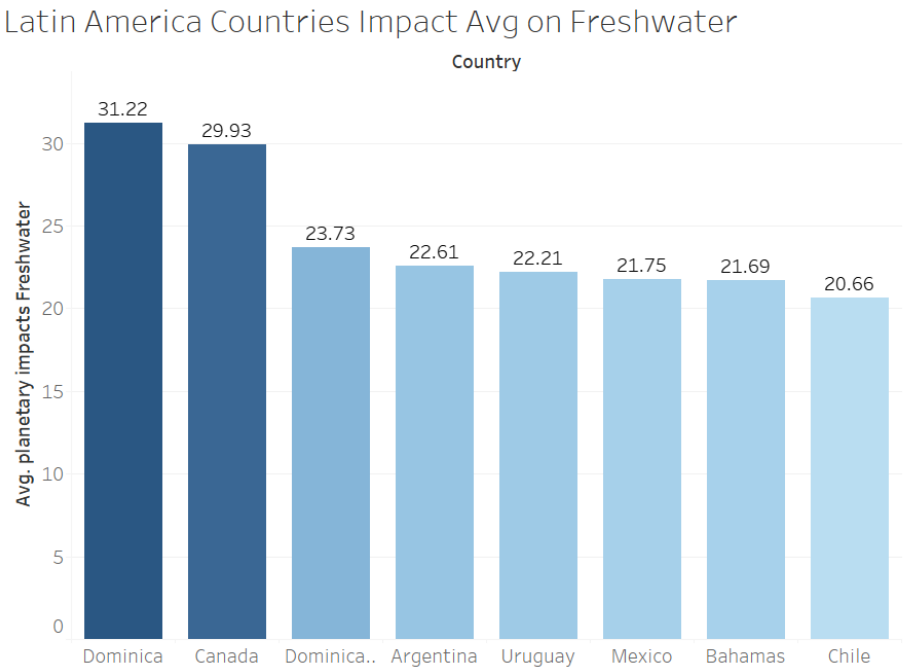


Figure 21: Bar graph that showcases Latin America countries with their average impact on freshwater

- In this figure it showcases that every Latin America country listed is above the average (18.34). This means that these countries have a poor use of their freshwater.

Dependency Analysis:

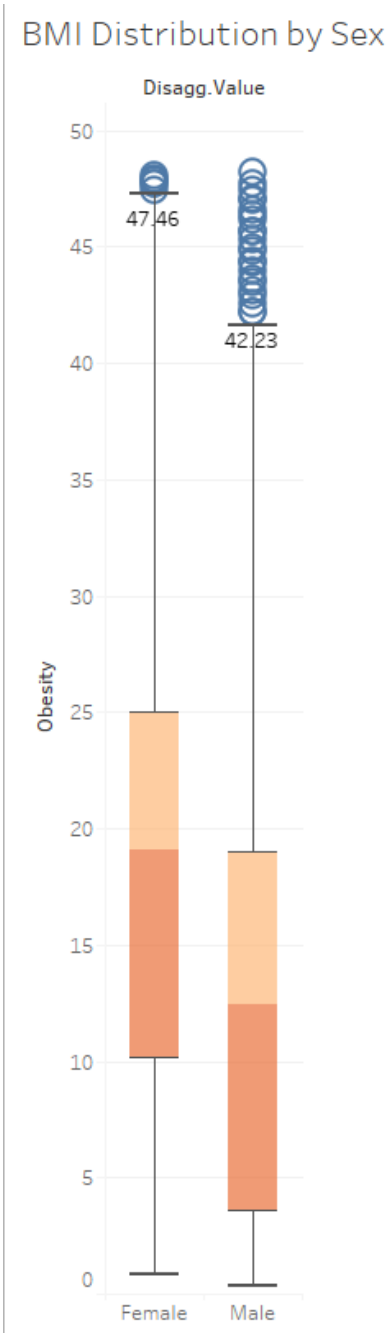


Figure 22: This box plot shows the spread of data between the BMI for females and males.

This figure illustrates the spread of data between males and females. Different sexes have different diets and food consumption recommendations; it is important to highlight the difference in obesity rates.

Predictive Models:

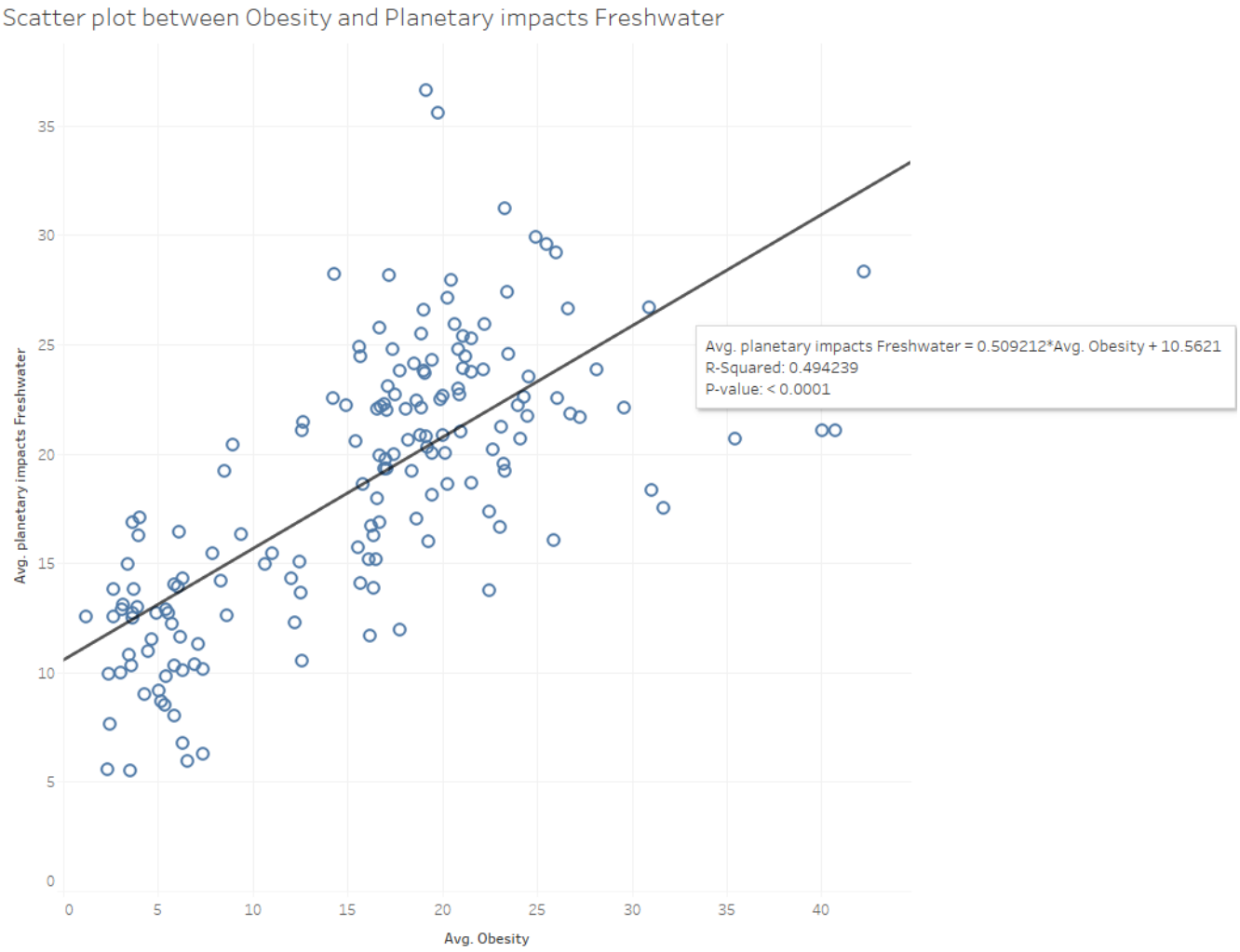


Figure 23: This scatterplot shows the correlation between obesity (calculated by BMI) and Food system impact on planetary boundary value (%) of freshwater.

- The food system impacts on planetary boundary of freshwater is the excess use of freshwater that causes consumable water scarcity, affecting agriculture and impacting the food production for local areas.
- In places where planetary impacts on freshwater are high and increases, the average obesity (BMI) increases.

Evaluation

Evaluate Results

The primary business objective of this project is to assist Dollar General in identifying global expansion opportunities by determining the most suitable countries within Asia and North & Latin America, focusing specifically on regions with nutritional deficiencies and environmental inefficiencies associated with high obesity rates. The goal is to recommend countries where Dollar General's low-cost essentials retail model can serve underserved populations.

Understanding and Interpreting Results

- The descriptive models collectively reveal a critical insight: regions with high obesity rates, specifically in Asia and North & Latin America, consistently show inadequate fruit and vegetable intake levels.
- Our analysis began with identifying what regions have the highest rates of obesity, Latin America and Asia (middle east) displayed the highest average in the data. We took 8 countries from Latin America and 10 from Asia
- The second part of our analysis was to identify what these countries had in common. We know fruits and vegetables are a crucial part of a healthy body, insufficient consumption contributes to unhealthy weight gain.
- In our analysis, we looked at the daily average fruit per gram intake for the countries we established earlier. We noticed they all fall below the recommended intake based on World Health Organization (WHO), revealing nutritional deficiencies.
- The predictive modeling scatter plot confirmed a strong, statistically significant relationship between freshwater impact and obesity rates ($R^2 = 0.494$, $p < 0.0001$), linking environmental factors with BMI.

Determine Next Steps

Possible Actions

- Deploy pilot/prototype stores in countries from Asia, North, and Latin America that exhibited high obesity rates and low fruit/vegetable intake
- Expand product line to include affordable and nutrient-dense foods
- Use environmental indicators, like freshwater usage, to determine which suppliers and products are most sustainable
- Introduce plant-based options and limit high processed foods
- Introduce market tags for products to highlight products with low-calories and low-processed foods
- Local government partnership to provide and educate locals on food health and obesity using Dollar General's product line.

Deployment

Analyzing Potential of each Action

- Deploy pilot/prototype stores in countries from Asia, North, and Latin America that exhibited high obesity rates and low fruit/vegetable intake: Highly feasible. Directly supports our business goal for Dollar General's international expansion
- Expand product line to include affordable and nutrient-dense foods: potential for deployment as it addresses Dollar General's affordable options while addressing the low nutritional intake issue.
- Use environmental indicators, like freshwater usage, to determine which suppliers and products are most sustainable: Strategically important but may require vast resources.
- Introduce plant-based options and limit high processed foods: semi feasible, the impact of planetary freshwater boundary affects agriculture and results in the relying on highly processed foods.

- Introduce market tags for products to highlight products with low-calories and low-processed foods: semi feasible, the locals can identify the healthier food options and make choices for themselves.
- Local government partnership to provide and educate locals on food health and obesity using Dollar General's product line: Unlikely. Every country has a different food culture, and it would be difficult to implement.

Ranking Possible Actions

1. Deploy pilot/prototype stores in countries from Asia, North, and Latin America that exhibited high obesity rates and low fruit/vegetable intake
2. Expand product line to include affordable and nutrient-dense foods
3. Introduce market tags for products to highlight products with low-calories and low-processed foods
4. Introduce plant-based options and limit high processed foods
5. Use environmental indicators, like freshwater usage, to determine which suppliers and products are most sustainable
6. Local government partnership to provide and educate locals on food health and obesity using Dollar General's product line

- The decision to deploy pilot or prototype stores in selected countries from Asia and Latin America is rooted directly in our project's business goal: identifying global expansion opportunities for Dollar General. This action offers the most immediate path to confirming market potential in regions with limited access to affordable and nutritious foods using high obesity rates and low fruit and vegetable intake.
- Summary of Deployable Results:
    - Identified target countries from North and Latin America & Asia with high obesity and low fruit/vegetable intake
    - Confirmed nutritional deficiencies using World Health Organization standards

- Determined environmental inefficiencies through predictive modeling
- Developed a clear list of possible business actions
- Ranked and determined best suitable action

Conclusions

This project successfully identified global expansion opportunities for Dollar General by analyzing countries in Asia, Latin America, and North America that have significant nutritional deficiencies and environmental inefficiencies, which are tied to rising obesity rates. By cleaning and describing our global nutrition report dataset, comprehensive modeling was created to discover critical information that supported the business goal. The final recommendation is to deploy pilot or prototype stores in the regions specified previously. This strategy supports Dollar General's long-term growth objectives and addresses their company goals.

Overall, this data-driven report provides a strong foundation for Dollar General's international growth strategy. Using this report, Dollar General can expand wisely and continue to serve their community internationally.

Glossary

Iso3: Unique country code
disagg.value: The unique value of each disaggregated value
Obesity: Adults aged 18 years and older with a BMI of 30 kg/m2 or higher (%)
Fruit: Estimated intake of fruit in adults aged 20 and older (g/day)
Vegetables: Estimated intake of vegetables in adults aged 20 and older (g/day)
Legumes: Estimated intake of legumes in adults aged 20 and older (g/day)
Nuts: Estimated intake of nuts in adults aged 20 and older (g/day)
Whole grains: Estimated intake of whole grains in adults aged 20 and older (g/day)
Fish: Estimated intake of fish in adults aged 20 and older (g/day)
Dairy: Estimated intake of dairy in adults aged 20 and older (g/day)

Red meat: Estimated intake of red meat in adults aged 20 and older (g/day)
Environmental_impacts_freshwater_use: Environmental footprint of food system components on freshwater use (km3)
Planetary_impacts_Freshwater: Food system impact on planetary boundary value (%) of freshwater
BMI: Body Mass Index

References

Huang, J., Ridoutt, B. G., Lan, K. (2020, January 2). ***Balancing food production within the Planetary Water Boundary*****. Journal of Cleaner Production.** https://www.sciencedirect.com/science/article/pii/S0959652619347705

Dataset and metadata**. 2021 Global Nutrition Report dataset and metadata - Global Nutrition Report. (n.d.).** https://globalnutritionreport.org/reports/2021-global-nutrition-report/dataset-and-metadata/

He, K., Hu, F. B., Colditz, G. A., Manson, J. E., Willett, W. C., & Liu, S. (2004, October 5). ***Changes in intake of fruits and vegetables in relation to risk of obesity and weight gain among middle-aged women*****. Nature News.** https://www.nature.com/articles/0802795#citeas

Nouri, M., Shateri, Z., & Faghih, S. (2023, January 17). ***The relationship between intake of fruits, vegetables and dairy products with overweight and obesity in a large sample in Iran: Findings of steps 2016*****. Frontiers.** https://www.frontiersin.org/journals/nutrition/articles/10.3389/fnut.2022.1082976/full

UNICEF/WHO/World Bank. Joint child malnutrition estimates expanded database: Stunting, wasting and overweight. https://data.unicef.org/resources/dataset/malnutrition-data

Tetens, I., & Alinia, S. (2015, November 7). The role of fruit consumption in the prevention of obesity. https://www.tandfonline.com/doi/abs/10.1080/14620316.2009.11512594

Tufts University. Global Dietary Database. Published online 2019. https://www.globaldietarydatabase.org/data-download