ANALYSIS OF FOOD CONSUMPTION PATTERNS ACROSS AMERICA

GROUP - 19

KULDEEP VISHAL CHOKSI
ARYAN PRAVIN AHER
ADHAVAN ALEXANDER

CONTENT

DATASET

The dataset utilized for this analysis was obtained from the official website of the Economic Research Service (ERS) of the U.S. Department of Agriculture, accessible via 'https://www.ers.usda.gov/data-products/food-consumption-and-nutrient-intakes/'. This dataset comprises a wide array of data points organized into multiple columns, meticulously curated to provide a comprehensive overview of food consumption and nutrient intake patterns prevalent across various demographics within the United States.

The primary objective of this research endeavor is to conduct a thorough and detailed analysis of the complex dynamics governing food consumption behaviors across America. Through rigorous examination and robust statistical methodologies, the goal is to identify discernible patterns, trends, and correlations within the dataset. These insights aim to offer valuable understanding into the dietary preferences, habits, and nutritional profiles of diverse populations across the nation. This analysis also seeks to explore the underlying factors influencing these consumption patterns, including socio-economic variables, cultural influences, and geographic disparities.

The dataset encompasses various columns written as follows:

1. "Food Group"
2. "Total_15_16"
3. "At home_15_16"
4. "Total_15_16_away"
5. "Restaurant_15_16_away"
6. "Fast food_15_16_away"
7. "School_15_16_away"
8. "Other_15_16_away"
9. "Total_17_18"
10. "At home_17_18"
11. "Total_17_18_away"
12. "Restaurant_17_18_away"
13. "Fast food_17_18_away"
14. "School_17_18_away"
15. "Other_17_18_away"

PREPROCESSING

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Food Group	Total_15_16	At home	_15_16_	rant_15_16	ood_15_16	ol_15_16_	_15_16_	otal_17_	home_17_	_17_18_	rant_17_18	ood_17_18	ol_17_18_	r_17_18_av
2	Added sugar														
3	Total population[1]	16.20	11.31	4.89	0.94	2.26	0.20	1.48	16.94	12.13	4.81	0.85	2.53	0.17	1.26
4	Children age 2–19	16.05	10.33	5.72	0.80	2.25	0.85	1.82	16.55	11.15	5.40	0.61	2.66	0.73	1.40
5	Adults age 20–64[2]	17.18	12.05	5.13	1.00	2.65	NA	1.49	17.76	12.68	5.07	0.96	2.89	NA	1.22
6	Seniors age 65 and above[2]	12.73	10.00	2.73	0.92	0.83	NA	0.98	14.37	11.49	2.87	0.79	0.93	NA	1.16
7	Household income < 185% poverty line	16.45	11.97	4.47	0.73	1.97	0.31	1.47	17.70	13.16	4.54	0.60	2.42	0.24	1.28
8	Household income 185–300% poverty line	17.01	11.56	5.45	1.05	2.38	0.18	1.84	18.94	14.41	4.53	0.82	2.55	0.22	0.94
9	Household income > 300% poverty line	15.60	10.56	5.04	1.09	2.48	0.12	1.35	15.49	10.33	5.16	1.09	2.63	0.09	1.35
10	Discretionary fats														
11	Total population[1]	34.86	22.82	12.04	2.95	5.87	0.55	2.67	36.24	23.92	12.32	2.64	6.66	0.45	2.58
12	Children age 2–19	33.73	21.30	12.43	1.96	5.28	2.29	2.89	33.84	21.95	11.89	1.31	6.00	1.88	2.71
13	Adults age 20–64[2]	36.00	22.95	13.04	3.42	6.96	NA	2.67	37.32	23.89	13.43	3.14	7.79	NA	2.51
14	Seniors age 65 and above[2]	32.33	24.63	7.69	2.68	2.69	NA	2.33	35.70	26.99	8.71	2.77	3.30	NA	2.64
15	Household income < 185% poverty line	33.74	22.93	10.80	2.25	5.11	0.87	2.57	34.11	22.65	11.46	1.76	6.34	0.67	2.70
16	Household income 185–300% poverty line	35.78	23.88	11.90	2.45	5.78	0.55	3.12	39.30	27.40	11.90	2.59	6.82	0.58	1.91
17	Household income > 300% poverty line	35.55	22.26	13.29	3.84	6.65	0.24	2.57	37.16	23.87	13.29	3.50	6.90	0.19	2.70
18	Discretionary oils														
19	Total population[1]	27.11	16.30	10.81	3.24	5.61	0.37	1.59	29.12	17.58	11.55	3.50	6.02	0.29	1.74
20	Children age 2–19	22.93	13.36	9.57	1.75	4.69	1.56	1.56	24.68	14.62	10.06	1.70	5.74	1.22	1.40
21	Adults age 20–64[2]	29.65	17.42	12.23	3.81	6.80	NA	1.61	31.36	18.30	13.06	4.17	7.03	NA	1.87
22	Seniors age 65 and above[2]	23.88	16.51	7.37	3.31	2.52	NA	1.54	27.22	19.26	7.95	3.61	2.59	NA	1.76
23	Household income < 185% poverty line	25.43	15.90	9.54	2.37	4.98	0.55	1.64	27.42	16.61	10.81	2.47	6.31	0.41	1.62
24	Household income 185–300% poverty line	25.76	15.88	9.88	3.09	5.09	0.36	1.34	28.15	17.85	10.29	2.87	5.92	0.33	1.17
25	Household income > 300% poverty line	29.32	16.88	12.44	4.14	6.44	0.21	1.65	31.09	18.40	12.69	4.69	5.79	0.16	2.05
26	Dairy														
27	Total population[1]	1.60	1.12	0.48	0.09	0.25	0.07	0.07	1.53	1.07	0.47	0.09	0.26	0.05	0.06
28	Children age 2–19	1.94	1.25	0.68	0.06	0.26	0.28	0.08	1.81	1.19	0.62	0.04	0.27	0.22	0.08
29	Adults age 20–64[2]	1.54	1.08	0.46	0.11	0.29	NA	0.06	1.47	1.00	0.47	0.10	0.30	NA	0.06
30	Seniors age 65 and above[2]	1.32	1.08	0.24	0.08	0.09	NA	0.07	1.34	1.12	0.22	0.08	0.09	NA	0.04
31	Household income < 185% poverty line	1.54	1.07	0.47	0.07	0.22	0.10	0.07	1.49	1.05	0.44	0.05	0.25	0.08	0.07
32	Household income 185–300% poverty line	1.67	1.20	0.47	0.08	0.24	0.07	0.08	1.58	1.15	0.43	0.08	0.23	0.07	0.05
33	Household income > 300% poverty line	1.63	1.13	0.50	0.12	0.28	0.03	0.06	1.55	1.05	0.50	0.13	0.28	0.02	0.07
34	Fruit														
35	Total population[1]	0.94	0.80	0.15	0.03	0.03	0.03	0.05	0.93	0.80	0.13	0.01	0.04	0.03	0.05
36	Children age 2–19	0.96	0.71	0.26	0.02	0.03	0.14	0.07	1.10	0.85	0.25	0.00	0.06	0.13	0.06
37	Adults age 20–64[2]	0.91	0.78	0.13	0.03	0.04	NA	0.05	0.83	0.73	0.10	0.02	0.04	NA	0.05
38	Seniors age 65 and above[2]	1.05	0.99	0.06	0.01	0.01	NA	0.04	1.07	1.00	0.08	0.02	0.01	NA	0.05
39	Household income < 185% poverty line	0.95	0.77	0.19	0.04	0.03	0.05	0.07	0.88	0.73	0.15	0.01	0.04	0.05	0.06
40	Household income 185–300% poverty line	0.85	0.71	0.13	0.01	0.03	0.03	0.05	0.94	0.82	0.12	0.01	0.04	0.03	0.04
41	Household income > 300% poverty line	0.98	0.86	0.11	0.02	0.04	0.01	0.04	0.97	0.85	0.12	0.01	0.04	0.01	0.05
42	Vegetables: total														
43	Total population[1]	1.40	0.92	0.47	0.16	0.22	0.02	0.08	1.39	0.90	0.49	0.17	0.23	0.01	0.08
44	Children age 2–19	0.90	0.56	0.33	0.06	0.14	0.07	0.06	0.86	0.53	0.33	0.06	0.18	0.05	0.04
45	Adults age 20–64[2]	1.57	1.02	0.55	0.19	0.27	NA	0.08	1.56	0.99	0.57	0.21	0.28	NA	0.09
46	Seniors age 65 and above[2]	1.49	1.10	0.39	0.19	0.10	NA	0.10	1.52	1.12	0.41	0.21	0.11	NA	0.09
47	Household income < 185% poverty line	1.26	0.86	0.40	0.11	0.17	0.03	0.08	1.24	0.79	0.45	0.12	0.22	0.02	0.09
48	Household income 185–300% poverty line	1.38	0.95	0.43	0.15	0.19	0.02	0.07	1.38	0.94	0.43	0.15	0.21	0.02	0.05
49	Household income > 300% poverty line	1.54	0.97	0.56	0.21	0.26	0.01	0.08	1.53	0.99	0.54	0.23	0.23	0.01	0.07

1. Data Reading: We use readxl to read the data file and convert it into our "data" format for preprocessing.

2. Numeric Conversion: Columns with numeric data initially read as characters due to mixed data types or 'NA' values are converted to numeric for mathematical operations.

3. Handling Missing Values: 'NA' strings representing missing values are converted to actual `NA` in R to ensure consistency and prevent misleading results.

4. Data Format Transformation: The dataset shifts from wide to long format for better analysis in R, especially for creating various plots.

5. Year and Location Extraction: Original column names containing year, location, and data type are separated for easy subsetting and faceting during visualization.

6. Aggregation: Data is aggregated to ensure each food group, year, and location combination has a single value for clarity and avoidance of duplications.

7. Data Cleaning: Column names are standardized by removing spaces or special characters, replaced with underscores for easier referencing and maintainable code.

PREPROCESSING

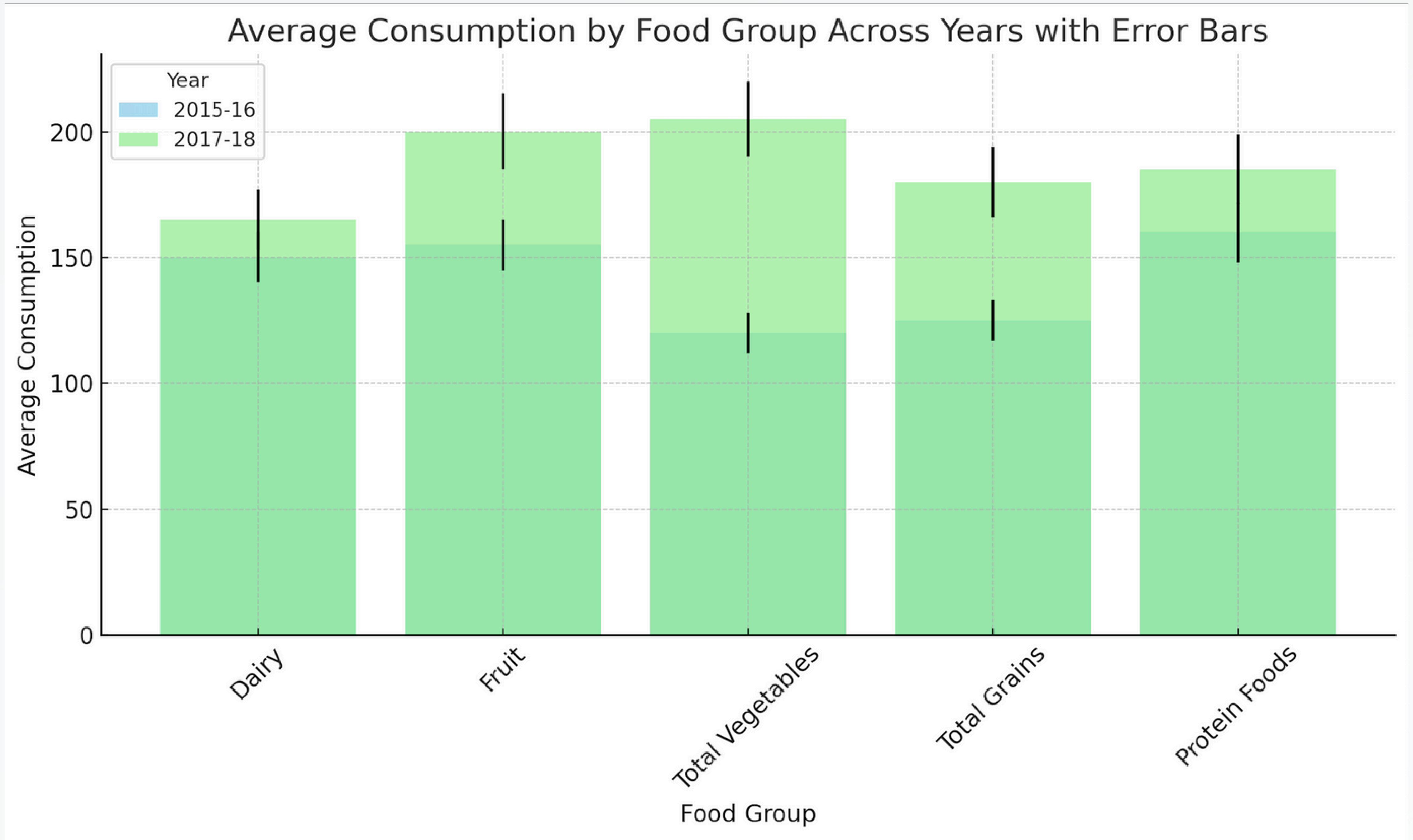
1. Preliminary Filtering and Cleaning:
   - We filtered the dataset to focus on essential food categories: Dairy, Fruit, Vegetables, Grains, and Protein Foods.
   - Using ‘select()’ and ‘filter()’, we isolated and retained relevant columns and rows for the years 2015-16 and 2017-18.
   - Rows with missing data (‘NA’ values) were removed using ‘drop_na()’ to ensure data accuracy.

2. Data Aggregation:
   - Data was grouped by food category using ‘group_by()’ for specific calculations.
   - ‘Summarise()’ computed average consumption for each food group across 2015-16 and 2017-18, excluding missing values with ‘na.rm = TRUE’.

3. Variability Assessment:
   - Using a similar approach, standard deviations were calculated for consumption data across both years, aiding in understanding consumption consistency across different food groups.

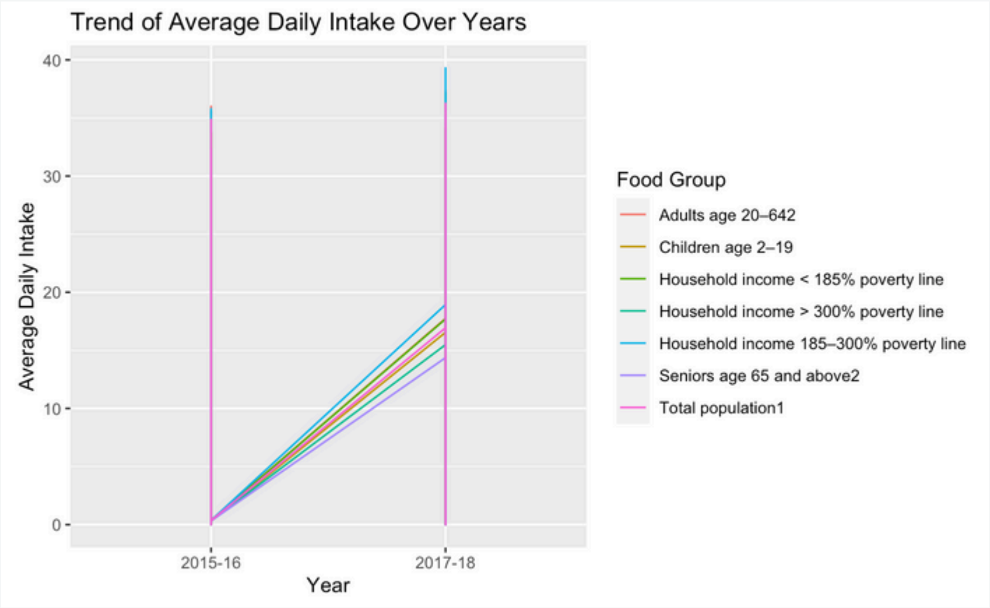
EXPLORATORY DATA ANALYSIS

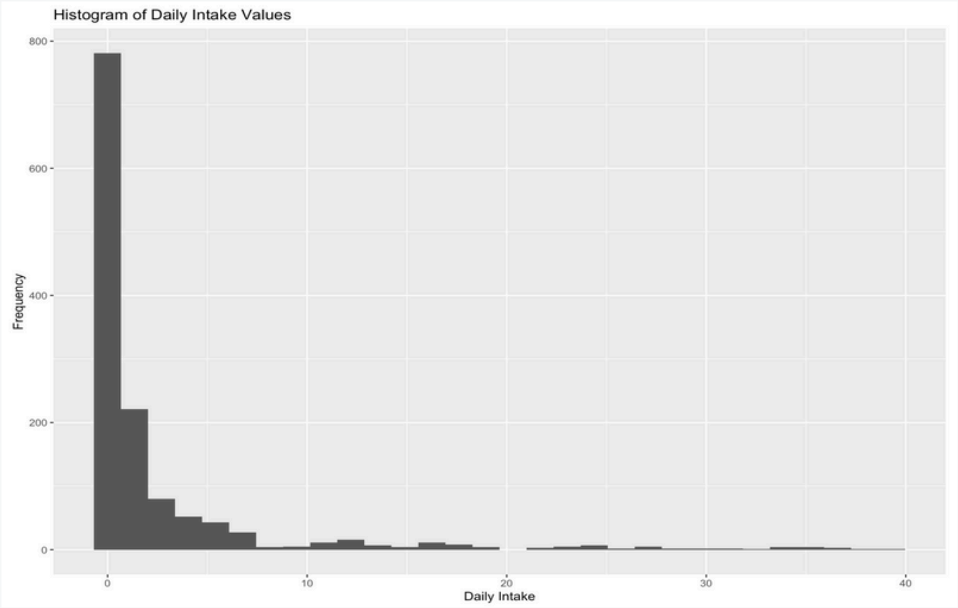
Thanks to the preliminary data processing, and as we aimed to obtain the following visualisation,

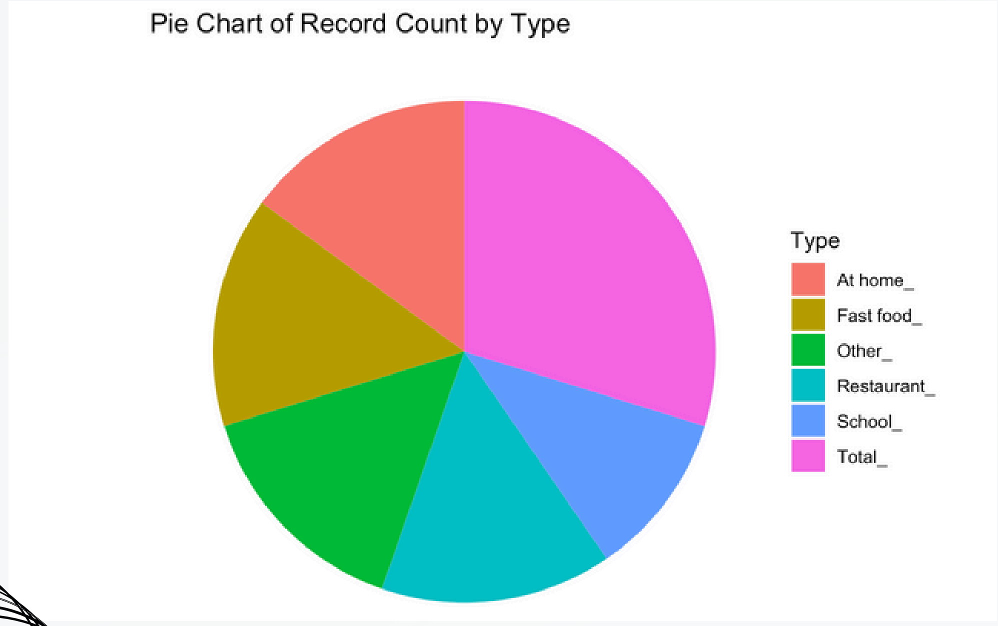


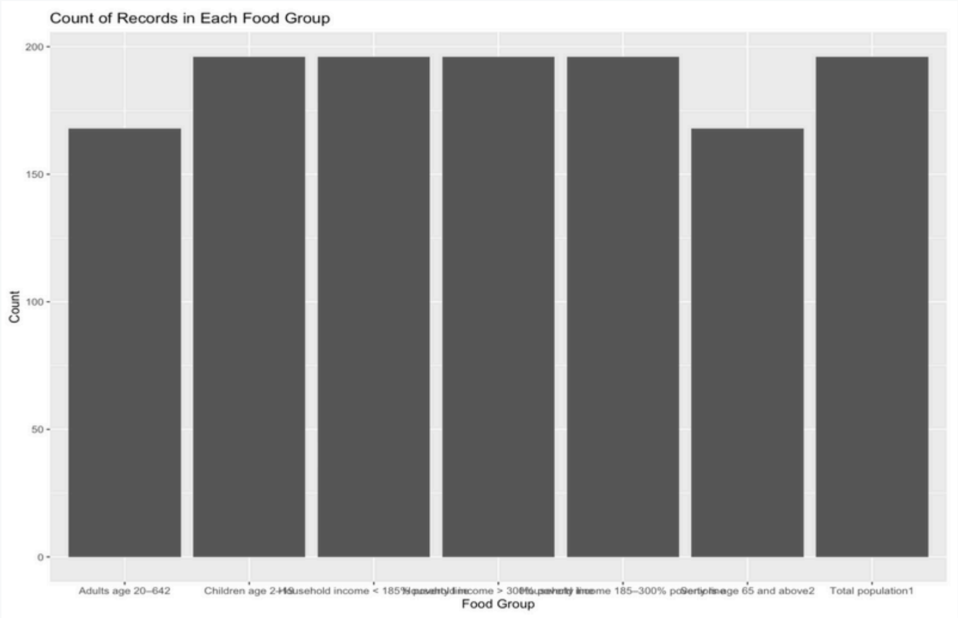
EXPLORATORY DATA ANALYSIS

However after going through multiple visualisations, as shown below, we obtained some final and proper visualisations that paint a pretty complex picture:



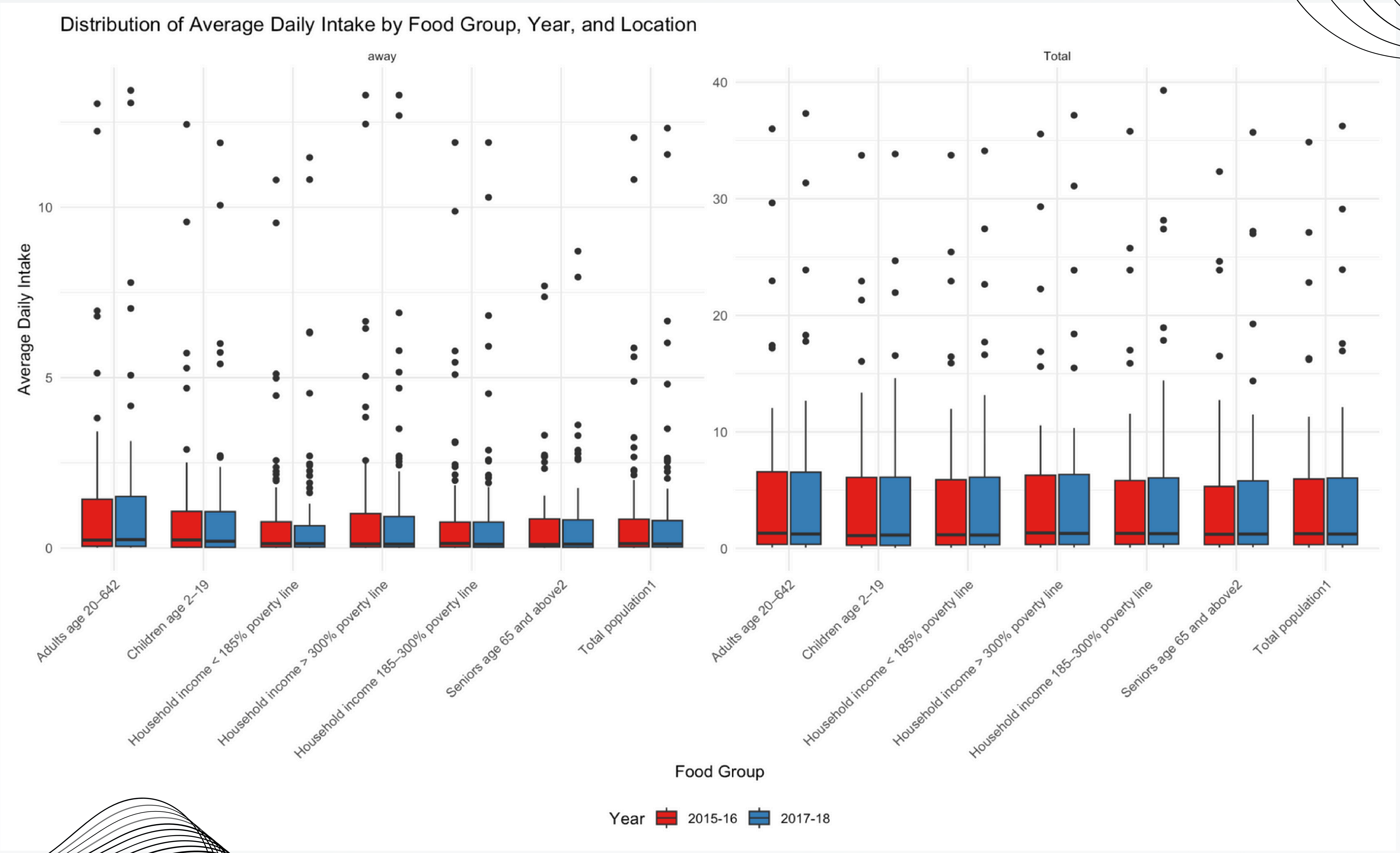






EXPLORATORY DATA ANALYSIS

EXPLORATORY DATA ANALYSIS



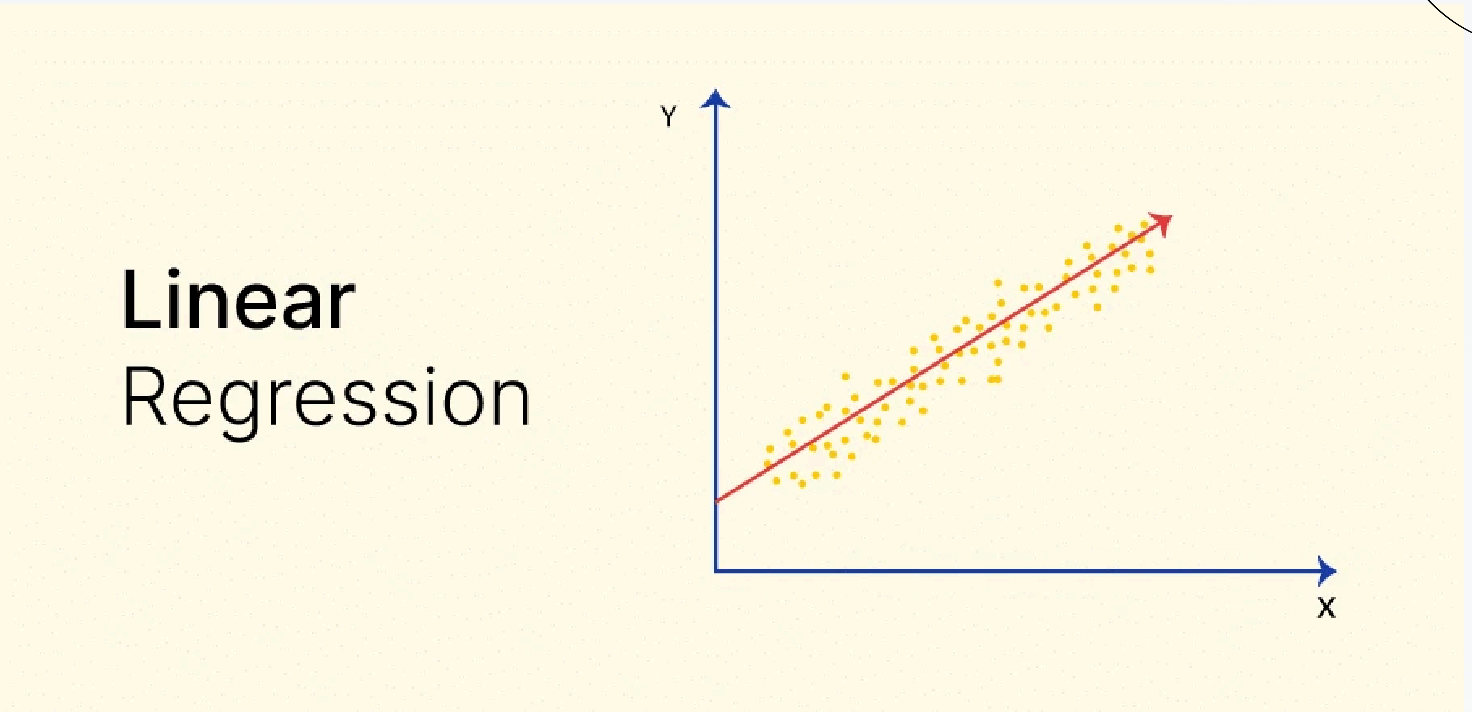
DATA MODELLING

In this analysis, we are interested in predicting the average daily intake of various food groups across different demographics. To achieve this, we've chosen 'Year' and 'Location' and 'type' as main predictors.
The target variable for our models is the 'value', representing the average daily intake. The goal is to understand the factors influencing this intake and to make informed predictions that could be useful for dietary planning or public health policy.

We compared several models to identify which would perform best for our predictive task:
- Linear Regression
- Random Forest
- Lasso Regression
- Ridge Regression
- Elastic Net
- Support Vector Regression (SVR)

DATA MODELLING, LINEAR REGRESSION

Linear regression is a statistical method to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables. The model aims to find the best-fit line that minimizes the difference between observed and predicted values. Coefficients represent the impact of independent variables on the dependent variable. The goal is to estimate these coefficients to make predictions about the dependent variable.

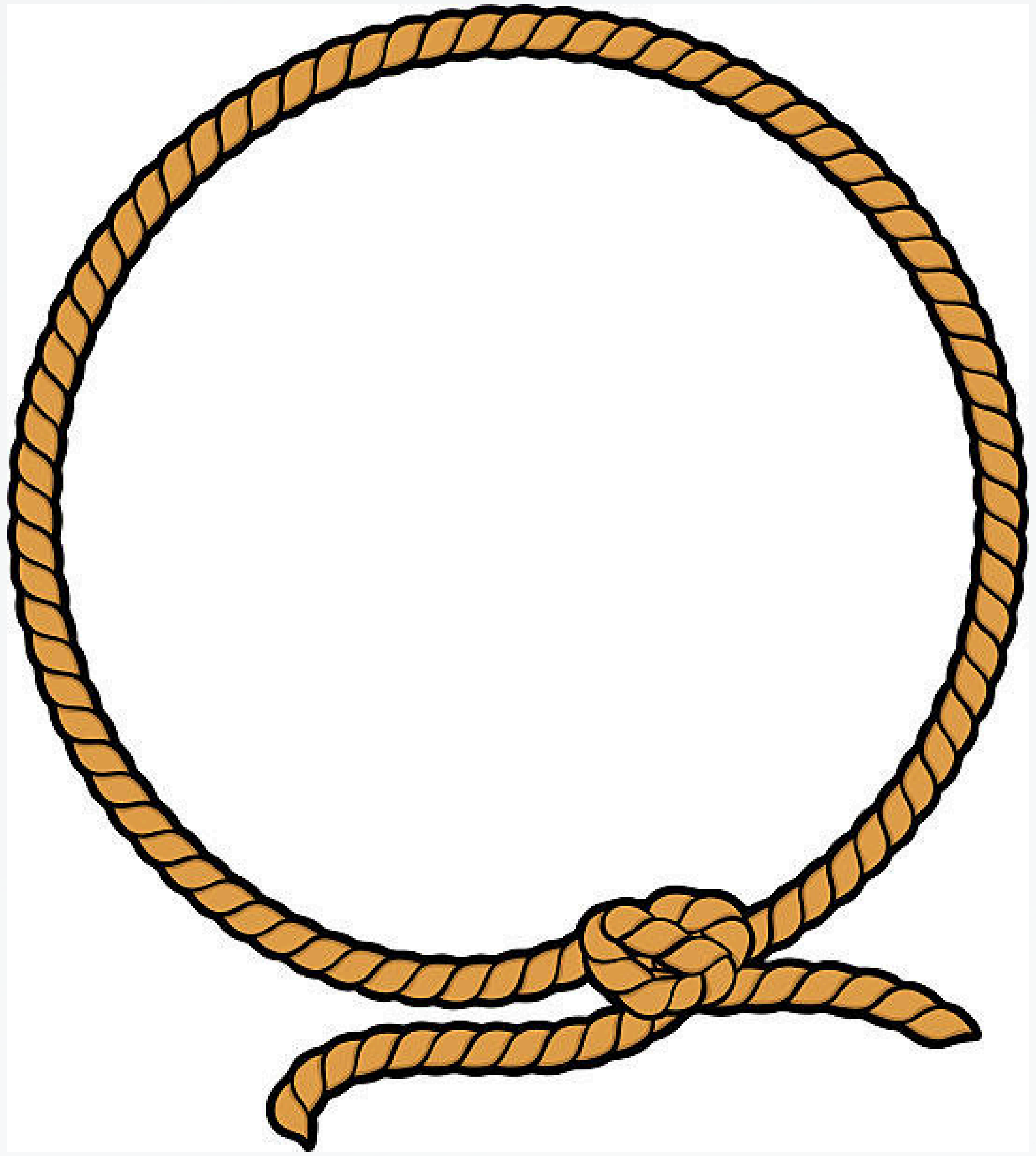


DATA MODELLING, RANDOM FOREST REGRESSION

Random Forest Regression is a powerful ensemble learning method that constructs a multitude of decision trees during training. Each tree is trained on a subset of the data and uses a random subset of features, introducing diversity. The final prediction is made by averaging the predictions of all individual trees, resulting in improved robustness and accuracy. Random Forests are less prone to overfitting compared to individual decision trees and can handle large datasets with high dimensionality effectively.

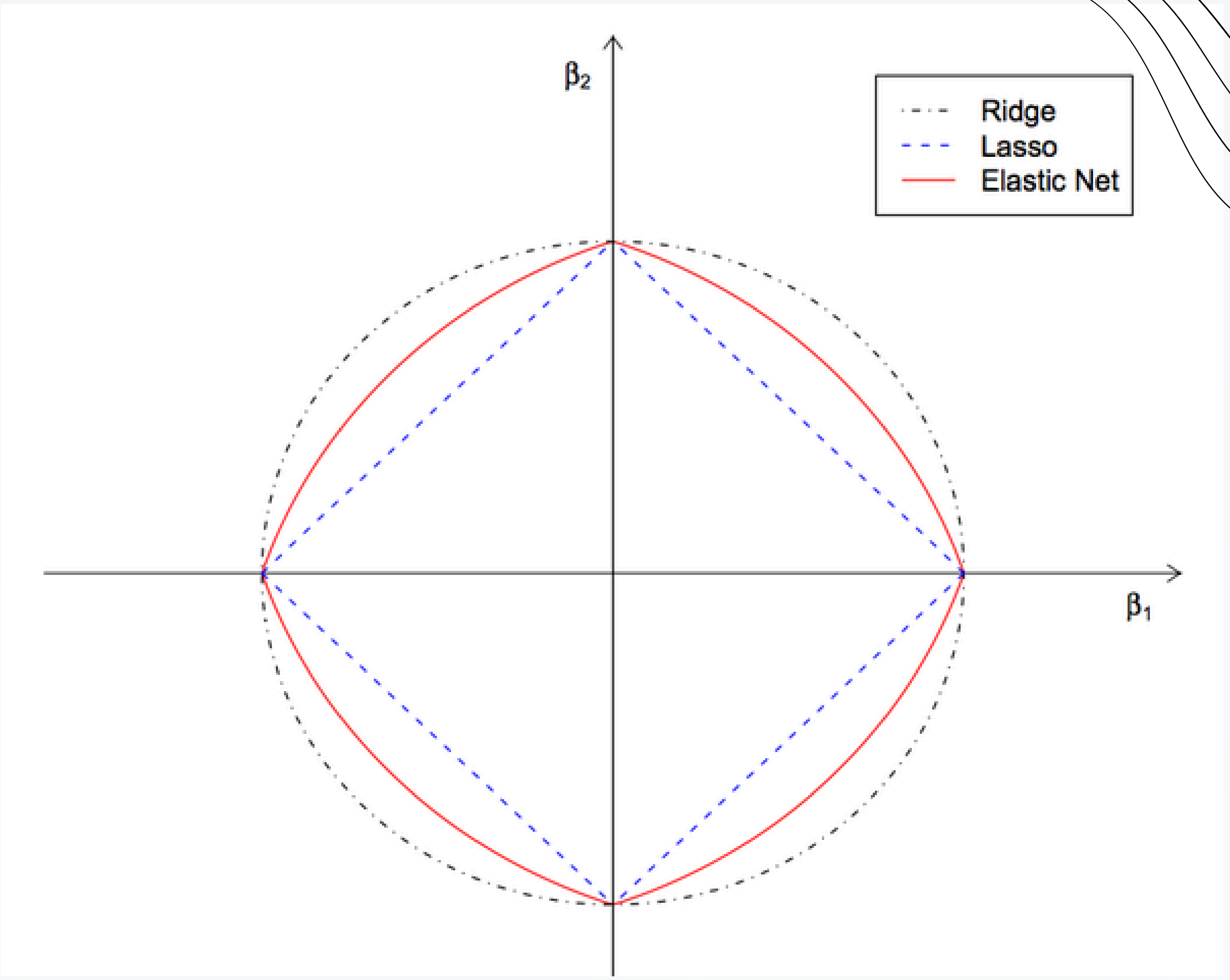
DATA MODELLING, LASSO

Lasso Regression, short for Least Absolute Shrinkage and Selection Operator, introduces an L1 penalty term to the ordinary least squares equation. This penalty term imposes a constraint on the sum of the absolute values of the coefficients. By penalizing large coefficients, Lasso encourages sparsity in the model, effectively performing feature selection by setting some coefficients to zero. It is particularly useful when dealing with datasets with a large number of features, as it can automatically select the most relevant ones and disregard the irrelevant ones.



DATA MODELLING, ELASTIC NET

Elastic Net is a regularization technique that combines the penalties of both Lasso and Ridge Regression. It introduces both L1 and L2 penalty terms to the ordinary least squares equation, providing a compromise between the advantages of Lasso and Ridge. Elastic Net is useful when dealing with datasets containing highly correlated features and when feature selection and regularization are both desired. By tuning the mixing parameter, Elastic Net allows for flexible control over the balance between L1 and L2 penalties.

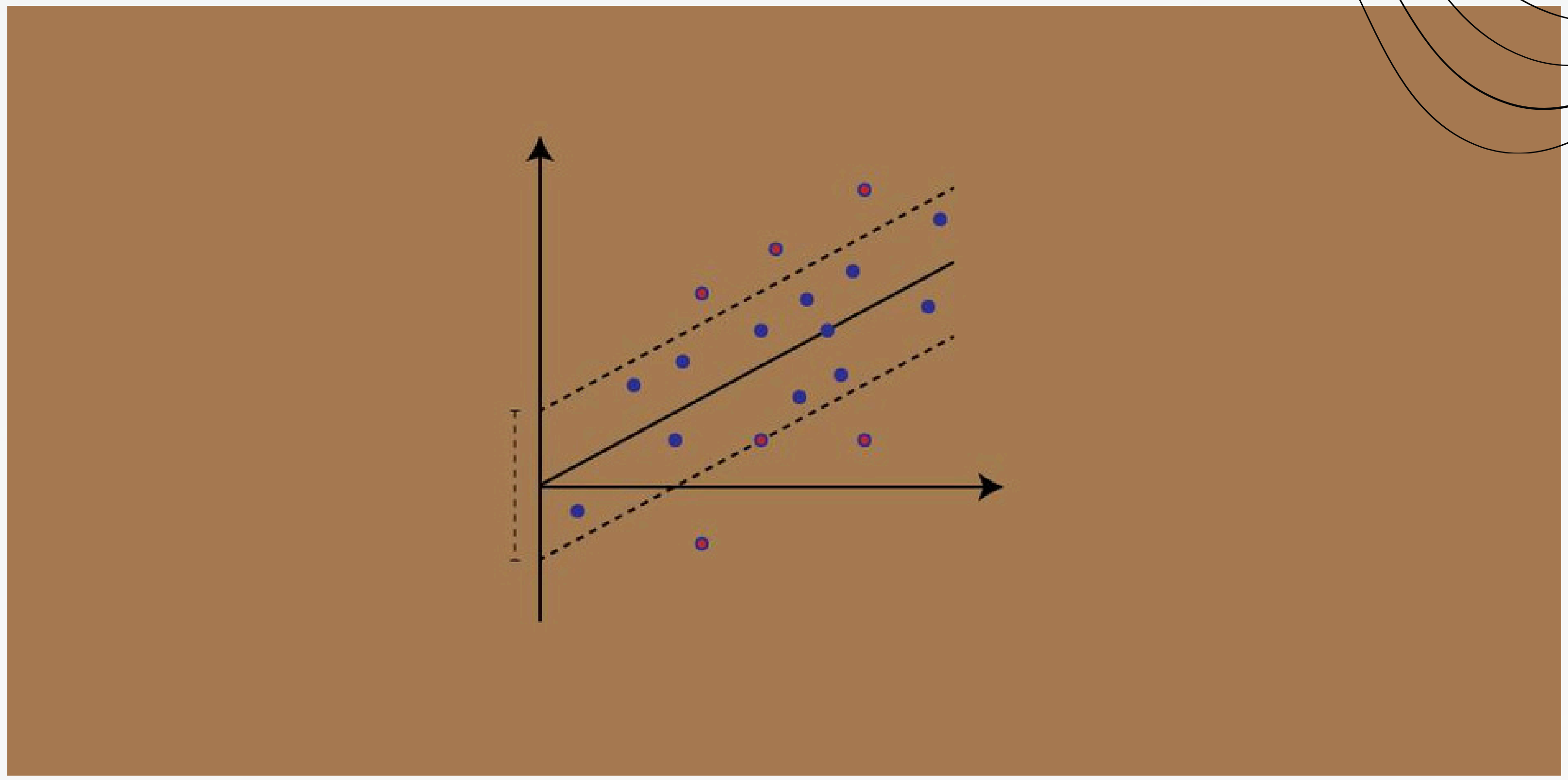


DATA MODELLING, RIDGE

Ridge Regression is a linear regression technique that introduces an L2 penalty term to the ordinary least squares equation. Unlike Lasso Regression, Ridge Regression penalizes the sum of squared coefficients. This penalty term helps prevent overfitting by shrinking the coefficients, effectively reducing their magnitudes. While Ridge Regression does not perform variable selection like Lasso, it can handle multicollinearity in the data by stabilizing the coefficients. It is particularly beneficial when dealing with datasets with highly correlated features.

DATA MODELLING, SVR

Support Vector Regression (SVR) is a regression algorithm based on the principles of Support Vector Machines (SVM). Unlike traditional regression techniques that aim to minimize error, SVR focuses on fitting a hyperplane within a specified margin of tolerance around the actual points. SVR is effective in capturing complex relationships in high-dimensional spaces and is particularly useful when dealing with datasets with nonlinear relationships. It can handle both linear and nonlinear regression tasks by using different kernel functions to map the data into a higher-dimensional space. SVR is robust to outliers and can generalize well to unseen data when properly tuned.



DATA MODELLING, COMPARISON

To compare the models, we split our dataset into training and testing sets, using 80% of the data for training and the remaining 20% for testing. The models were then trained on the training set, and predictions were made on the test set.

For each model, we calculated:

- Root Mean Squared Error (RMSE): Provides a measure of the model's prediction error in the same units as the target variable.
- Mean Absolute Error (MAE): Indicates the average absolute error in the model's predictions, offering a clear measure of average prediction error magnitude.
- R-squared (R^2): Shows the proportion of variance in the target variable that is predictable from the predictors, giving us an idea of the model's explanatory power.

RESULTS

The results were close across all models, with Lasso and Elastic Net showing a marginally better performance in terms of RMSE and R^2, suggesting a slightly better fit and prediction capability. The SVR model, although not the best in RMSE or R^2, had the lowest MAE, indicating its predictions were closer to the actual values on average. **The Elastic Net model** has yielded some insightful results

	RMSE <dbl>	MAE <dbl>	R2 <dbl>
Linear Model	5.065619	2.762874	0.1412189
Random Forest	5.074257	2.794262	0.1399284
Lasso	4.983220	2.660401	0.1682801
Ridge	4.983912	2.647408	0.1685622
Elastic Net	4.983220	2.660401	0.1682801
SVR	5.494307	2.276713	0.1571687

<u>Balance Between Overfitting and Underfitting:</u> Elastic Net combines L1 and L2 regularization, which can handle collinear features better and is less likely to overfit than a simple linear regression model.

<u>Predictive Power:</u> Even though the R^2 value might not be high, the model is still providing predictive power above the baseline, which could be valuable in certain contexts or as a starting point for further refinement.

RESULTS

Upon modelling we can have an idea about how thr predicted values are when compared to the original values, we use the head function to get a tabular format of it.

A tibble: 6 × 7

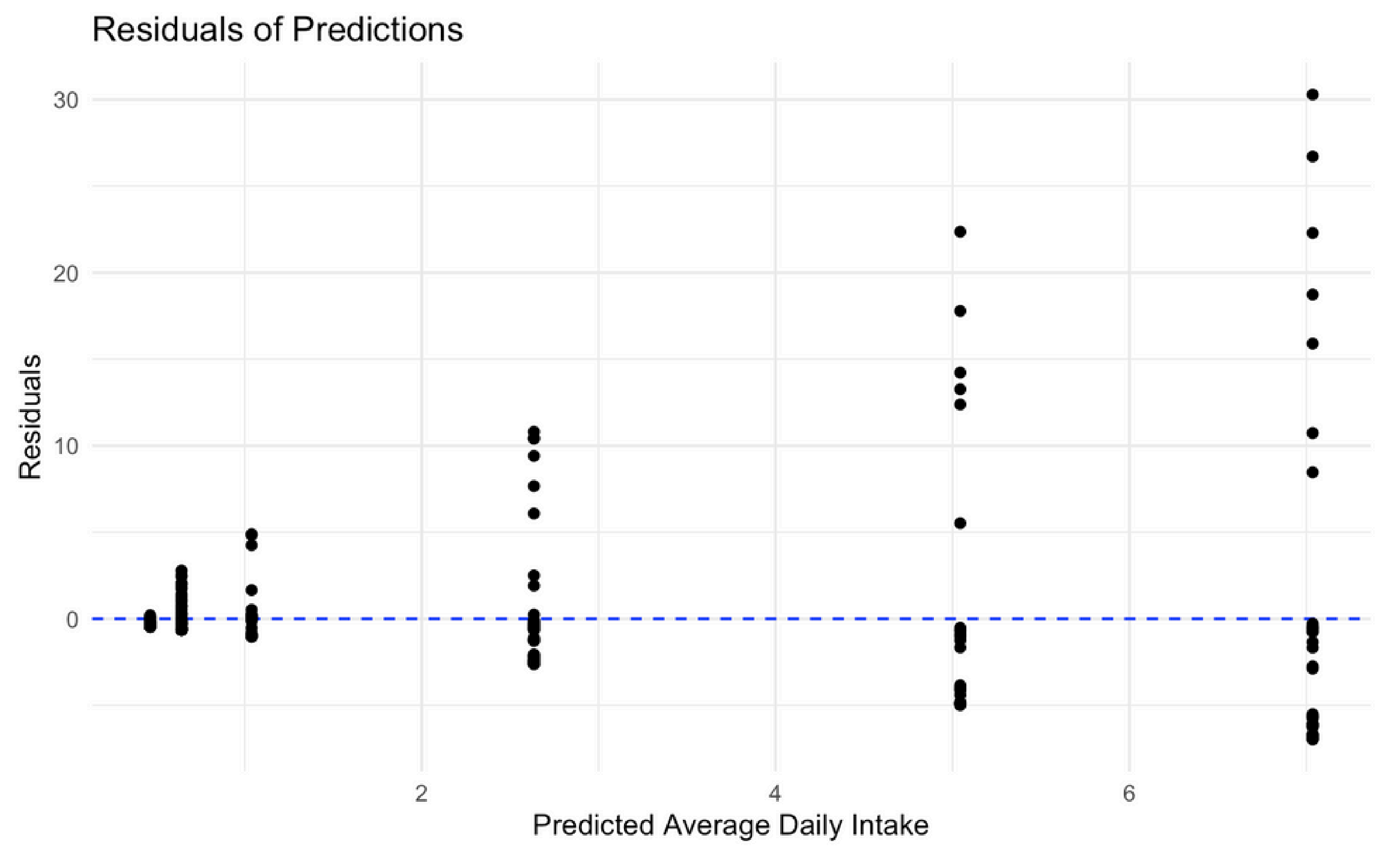
Food Group <chr>	Type <chr>	Year <fctr>	Location <fctr>	value <dbl>	predicted_value <dbl>	residuals <dbl>
Total population1	Restaurant_	2015-16	away	0.94	0.6427766	0.2972234
Adults age 20-642	Total_	2015-16	away	5.13	2.6346658	2.4953342
Adults age 20-642	Total_	2017-18	Total	17.76	7.0352604	10.7247396
Adults age 20-642	Restaurant_	2017-18	away	0.96	0.6427766	0.3172234
Seniors age 65 and above2	Total_	2017-18	away	2.87	2.6346658	0.2353342
Household income < 185% poverty line	School_	2015-16	away	0.31	0.4654552	-0.1554552

6 rows

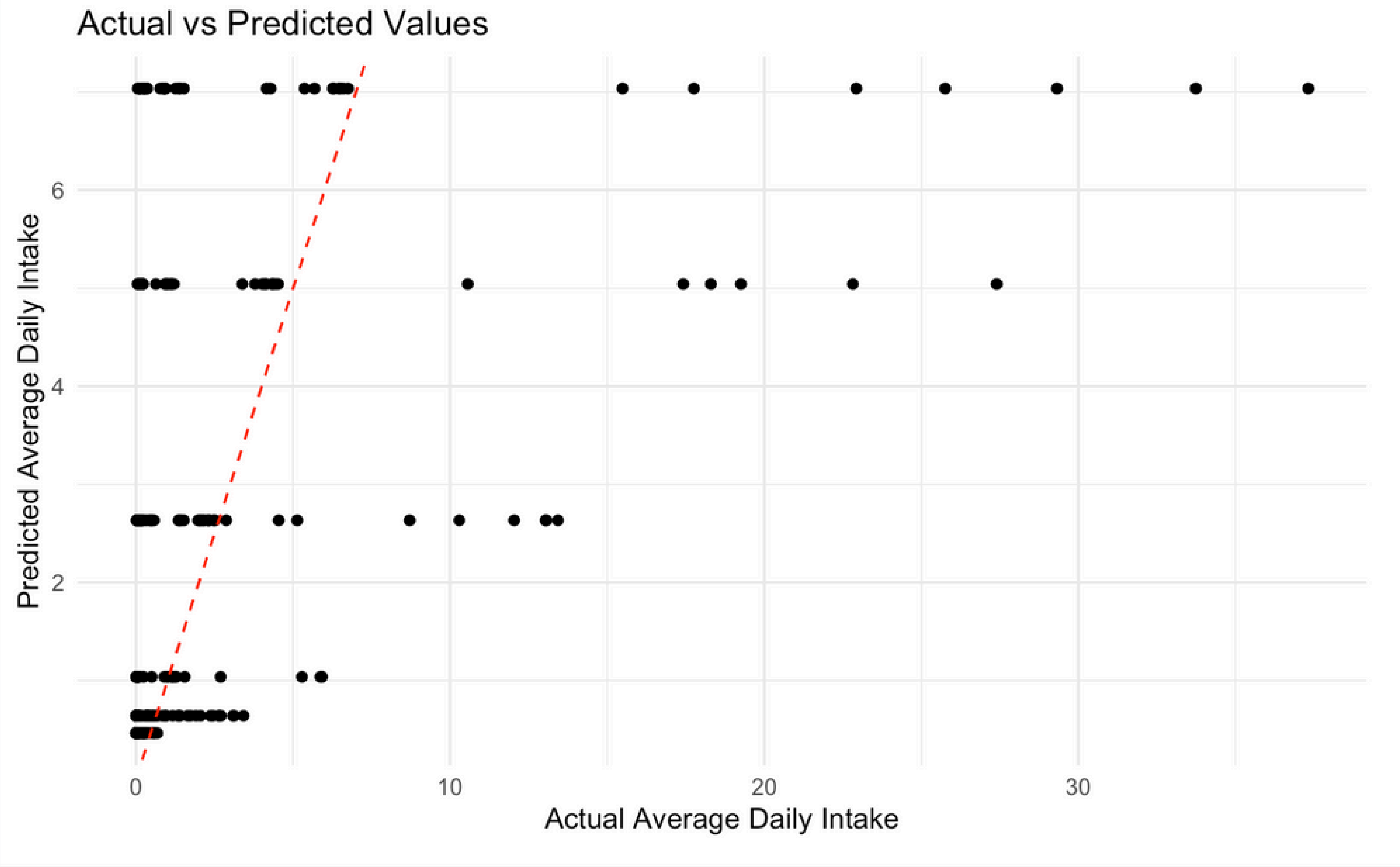
RESULTS

1. Predicted values:
   - The predicted values range from around 0.2 (for "Household income < 185% poverty line" with type "School_" in 2017-18) to around 17.8 (for "Adults age 20-642" with type "Total_" in 2017-18).
   - Higher predicted values generally correspond to food groups with larger populations or higher income levels, such as "Total population1" and "Household income > 300% poverty line."
   - Lower predicted values are associated with more specific or narrower segments, like "Seniors age 65 and above2" or lower-income groups.
2. Residuals:
   - Residuals represent the difference between the actual observed value and the predicted value. Positive residuals indicate that the model underpredicted the value, while negative residuals indicate overprediction.
   - Some large positive residuals can be seen for rows like "Total population1" with type "At home_" in 2015-16 (residual = 17.776628819) and "Household income < 185% poverty line" with type "Total_" in 2015-16 (residual = 26.704739633). This suggests that the model significantly underpredicted the values for these groups.
   - Conversely, there are large negative residuals for rows like "Adults age 20-642" with type "Total_" in 2017-18 (residual = -30.284739633) and "Adults age 20-642" with type "Total_" in 2015-16 (residual = -27.767739633), indicating that the model overpredicted the values for these groups.

RESULTS



Residuals of Predictions
Residuals
Predicted Average Daily Intake
0
10
20
30
2
4
6



Actual vs Predicted Values
Predicted Average Daily Intake
Actual Average Daily Intake
2
4
6
0
10
20
30

LIMITATIONS

1. Predicting Extreme Values: The "Actual vs Predicted Values" plot shows that the model does not capture the extreme values well, as most predictions are concentrated around the lower end of the scale.
2. Residual Patterns: In the "Residuals of Predictions" plot, the residuals should ideally be scattered randomly around the horizontal line at 0. Instead, there's a visible pattern, suggesting that the model might not be capturing some of the complexity in the data.
3. Room for Improvement: The actual vs. predicted plot doesn't show a tight alignment along the diagonal line, which indicates that there is still room for model improvement. This might involve feature engineering, model tuning, or trying different modeling techniques.
4. Low R^2 Value: The low R^2 value indicates that the model's predictions are not highly correlated with the actual values, which means the model doesn't explain a large portion of the variance in the response variable.

FUTURE SCOPE

In terms of overall assessment, while the Elastic Net model demonstrates some predictive capabilities, particularly in terms of controlling for overfitting, it may require additional work to improve its prediction of higher values and to increase the overall R^2. Potential steps to improve the model could include incorporating additional relevant features, transforming existing features to better capture non-linear relationships, or tuning model hyperparameters.

THE END

REFERENCES

- https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.voxco.com%2Fblog%2Fhow-can-you-calculate-linear-regression%2F&psig=AOvVaw23fhT18YjcKdaAQtegyo28&ust=1713406935440000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCKj5sK-YyIUDFQAAAAAdAAAAABAE
- https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.analytixlabs.co.in%2Fblog%2Frandom-forest-regression%2F&psig=AOvVaw2g_dGDcPhq5RI2xEddiurT&ust=1713407089814000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCID7tPmYyIUDFQAAAAAdAAAAABAE
- https://www.google.com/imgres?q=lasso&imgurl=https%3A%2F%2Fmedia.istockphoto.com%2Fid%2F480202786%2Fvector%2Frope-border-lasso.jpg%3Fs%3D612x612%26w%3D0%26k%3D20%26c%3DIuc-V32WrQDkPC8UsNy1bwqhKZA5pAUottlWRcvpK3k%3D&imgrefurl=https%3A%2F%2Fwww.istockphoto.com%2Fillustrations%2Flasso-cowboy-rope-cartoon&docid=9RYli_lX78LHCM&tbnid=E3RFsXNtbzJLmM&vet=12ahUKEwit14-FmciFAxUmNlkFHWL3Ca4QM3oECGUQAA..i&w=544&h=612&hcb=2&ved=2ahUKEwit14-FmciFAxUmNlkFHWL3Ca4QM3oECGUQAA
- https://www.google.com/imgres?q=elastic%20net&imgurl=https%3A%2F%2Fcdn.corporatefinanceinstitute.com%2Fassets%2Felastic-net1-1024x642.png&imgrefurl=https%3A%2F%2Fcorporatefinanceinstitute.com%2Fresources%2Fdata-science%2Felastic-net%2F&docid=4hfqk5aQtBcrAM&tbnid=5sN8zQZJjNgRnM&vet=12ahUKEwj-o6CvmciFAxUiEVkFHUZABioQM3oECBQQAA..i&w=1024&h=642&hcb=2&ved=2ahUKEwj-o6CvmciFAxUiEVkFHUZABioQM3oECBQQAA
- https://www.google.com/url?sa=i&url=https%3A%2F%2Fohwhataknight.co.uk%2Fblog%2FNantlle-ridge-hike-guide&psig=AOvVaw3q4lSUFYzGjr0H9sDe480v&ust=1713407287080000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCNCa49eZyIUDFQAAAAAdAAAAABAE
- https://www.google.com/imgres?q=svr%20modelling&imgurl=https%3A%2F%2Fcdn.analyticsvidhya.com%2Fwp-content%2Fuploads%2F2020%2F03%2FSupport-Vector-Regression.gif&imgrefurl=https%3A%2F%2Fwww.analyticsvidhya.com%2Fblog%2F2020%2F03%2Fsupport-vector-regression-tutorial-for-machine-learning%2F&docid=LlbCRNkyRExwbM&tbnid=FP2n7_BGWCkafM&vet=12ahUKEwiM5IrzmciFAxXEFVkFHVpBAwUQM3oECHUQAA..i&w=1050&h=520&hcb=2&ved=2ahUKEwiM5IrzmciFAxXEFVkFHVpBAwUQM3oECHUQAA