01.020 DESIGN THINKING PROJECT III

Modelling Uncertainty

Ecolet

Predictive Food Waste Model

Composed By SC11 Group 4:

Davin Handreas Chen	1009595
Thong Zi Qi	1009160
Catherine Laura Danandjaja Antoni	1009263
Ryan Leow Heng Kuan	1009273
Wong Jun Han Brayden	1009259

Background

Food waste is a significant global challenge with severe environmental, economic, and social consequences. Singapore alone wastes over $342 million worth of food annually (Singapore Environment Council, 2019). This substantial waste stream burdens landfills, emits greenhouse gases, and represents a considerable loss of resources.

Existing digital solutions aimed at tackling this problem are often limited. Many current tools are designed for post-hoc waste tracking, meaning they help businesses record waste after it has already occurred. While useful for auditing, this approach does not provide the proactive insights needed to prevent waste from happening in the first place. Instead of only reacting to food waste, this project introduces a predictive model that helps businesses anticipate and minimize food wastage before it occurs.

Problem Statement

How might we help F&B businesses in Singapore forecast food waste before it happens, enabling smarter planning and more sustainable procurement?

Purpose of Model

This model aims to allow F&B owners to:

- Shift food waste management from reactive tracking to proactive planning.
- Make **informed decisions** on purchasing and preparation quantities, avoiding overproduction.
- Support sustainable practices within the catering industry by enabling data-driven forecasting and resource optimisation.

Data Set Acquisition

The dataset for this project, "*Food wastage data in restaurant*" provides a comprehensive collection of restaurant-specific data in India, designed to offer insights into the patterns and magnitude of food waste. The dataset has been curated to ensure the anonymity and confidentiality of the restaurant and its patrons, making it suitable for academic and research purposes.

Variables

a. Independent (Predictor) Variables

These variables are the selected predictors that will be used to explain the variation in the amount of food wastage.

Table 1. Independent Variables Classification

Variable	Type	Description
Type of Food	Categorical	The category of food served (e.g., Meat, Vegetables, Fruits, Dairy Products, Baked Goods)
Pricing	Categorical	The price level of the meal or event (e.g., Low, Moderate, High)
Event Types	Categorical	The type of service or event (e.g., Birthday, Corporate, Wedding, Social Gathering)
Geographical Location	Categorical	Restaurant Location (e.g., Urban, Suburban, Rural)
Preparation Method	Categorical	The way of how food is served (e.g., Buffet, Sit-down Dinner, Finger Food)
Number of Guests	Numerical	The number of patrons served.

Quantity of Food (kg)	Numerical	The total quantity of food prepared or served.

b. Dependent (Response) Variables

The response variable is the variable that the model aims to predict, which will be the **amount of food wastage** in kilograms. The multiple linear regression model will use the independent variables to predict this value.

Data Cleaning

Categorical variables were converted into dummy variables for regression. For each categorical feature, one category was intentionally excluded (*n-1* variables) and used as the baseline to prevent perfect multicollinearity. The coefficients of the included dummy variables then represent the effect of each category relative to the baseline, which enhances interpretability and preserves model validity (Loy, 2015). **(Refer to APPENDIX I: *Dummy Variables & Baselines*)**

Additionally, to ensure the effectiveness of our multiple linear regression model, we first examined the correlation matrix of all variables, including both numerical and dummy-encoded categorical variables. This matrix helps to identify multicollinearity, check linear independence, and spot redundancy of variables. High multicollinearity can distort the estimation of regression coefficients, making it difficult to determine the individual effect of each predictor.

Based on the correlation matrix shown in *Appendix II*, the majority of predictor variables exhibit low to moderate correlation coefficients, indicating that they do not have strong linear relationships with one another. One exception is the strong positive correlation observed between *Number of Guests* and *Quantity of Food*, which is expected given their conceptual link. To avoid potential multicollinearity in the initial model, *Number of Guests* was initially excluded, but later reintroduced after evaluating its contribution to model performance. **(Refer to APPENDIX II: *Correlation matrix for predictor variables*)**

Assessment of Categorical Predictors via Boxplot Analysis

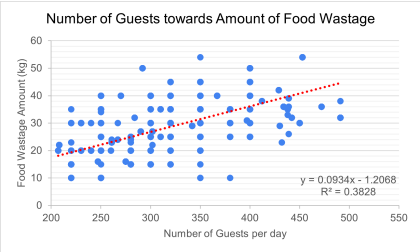
Boxplots were used to assess the distribution of food wastage across five categorical variables: *Type of Food, Pricing, Type of Event, Geographical Location*, and *Preparation Method*. The categories appear well balanced with no extreme skewness or underrepresented groups. Additionally, the relatively consistent spreads across categories suggest homoscedasticity, which is the variance of wastage remains stable across different groups, further supporting the assumptions of linear regression. **(Refer to APPENDIX III: *Box Plot Analysis of Categorical Datas*)**

Preliminary Linear Regression for Numerical Variables

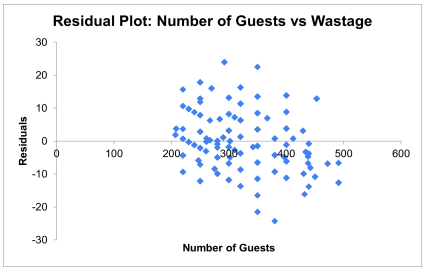
Before constructing the multilinear regression model, a preliminary linear regression was conducted using *Number of Guests* and *Quantity of Food* as predictor variables against the *Food Wastage Amount*. This is to ensure that the two numerical predictor variables are statistically significant and relevant as key predictors.

a. Number of Guests x Food Wastage Amount

Graph 1. Regression of Number of Guests vs. Food Wastage



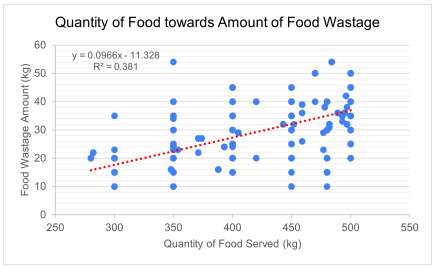
Graph 2. Residual Plot of Number of Guests vs.Food Wastage



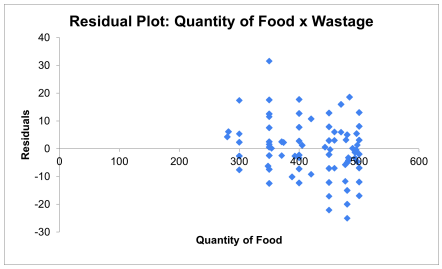
Based on the linear regression model, we can interpret that for every additional guest, food wastage is expected to increase by approximately 0.0934 kilograms. Based on *Appendix IV*, we obtain an adjusted R^2 value of 0.3828, indicating that the number of guests accounts for a moderate portion of the variation in food wastage. Additionally, the p-value for the slope coefficient is less than alpha value 0.05, making the model statistically significant. The residuals are also fairly randomly scattered around the horizontal axis with no clear pattern, which indicates that the linear model is appropriate and that the assumption of linearity is reasonably satisfied. **(Refer to APPENDIX IV: *Data Analysis for No. of Guests vs. Food Wastage Regression)***

b. Quantity of Food x Food Wastage Amount

Graph 3. Regression of Quantity of Food vs. Food Wastage



Graph 4. Residual Plot of Quantity of Food vs. Food Wastage



The linear regression analysis indicates a statistically significant positive relationship between the quantity of food and food wastage. Specifically, for every additional food in kilograms, food wastage is predicted to increase by 0.0966 kilograms. The model, with an adjusted R^2 of 0.3810, suggests that the quantity of food explains approximately 38.1% of the variation in food waste. The residuals again appear to be randomly distributed across the x-axis, with no strong patterns, curvature, or clustering. This supports the assumption of homoscedasticity and linearity in the relationship between quantity of food and food wastage. **(Refer to APPENDIX V: *Data Analysis for Quantity of Food vs. Food Wastage Regression)***

Multilinear Regression - 1st Iteration

In the initial model, all variables were included as predictors of *Food Wastage Amount*. This model yielded an adjusted R^2 of 0.749, indicating a relatively strong fit. The standard error of the regression is 5.05, which indicates that the predicted food wastage values deviate from the actual values by approximately ±5.05 kilograms on average. This provides an estimate of the model's overall prediction accuracy. Additionally, the model is based on a dataset comprising 1,609 observations, which represents a robust sample size. Such a large sample contributes to the statistical reliability and generalizability of the regression results.

However, upon closer examination of the regression output, several predictors namely *Type of Food*, *Event Type*, and *Geographical Location*, all exhibited p-values greater than 0.05, suggesting that their effect on the dependent variable was not statistically significant at the 95% confidence level. Including such non-significant variables can increase the complexity of the model unnecessarily and lead to overfitting, where the model captures noise rather than meaningful patterns. Overfitted models may perform well on training data but tend to generalize poorly to new, unseen data. Therefore, even though the model performs decently, keeping variables with high p-values may dilute the explanatory power and reduce overall interpretability. Removing them would likely lead to a more accurate model, which leads to the 2nd iteration of our model. **(Refer to APPENDIX VI: *1st iteration of multilinear reg. model)***

Multilinear Regression - Final Iteration

Table 2. Correlation Matrix for Predictor Variables

Regression Statistics								
Multiple R	0.87586995							
R Square	0.76714817							
Adjusted R Square	0.76627606							
Standard Error	4.88447694							
Observations	1609							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	6	125921.0202	20986.84	879.6519235	0			
Residual	1602	38220.70014	23.85811					
Total	1608	164141.7203						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	16.5760132	1.013097725	16.36171	9.21098E-56	14.58887677	18.5631495	14.58887677	18.56314954
PreparationMethod_Buffet	-1.00938355	0.352234508	-2.86566	0.004215762	-1.700272484	-0.3184946	-1.70027248	-0.318494619
PreparationMethod_FingerFood	-4.31884232	0.294922333	-14.644	1.14502E-45	-4.897316526	-3.7403681	-4.89731653	-3.740368122
Pricing_Low	-12.9772008	0.328361997	-39.521	5.1003E-239	-13.62126509	-12.333137	-13.6212651	-12.3331365
Pricing_Moderate	-11.2559534	0.31231726	-36.0401	8.3318E-209	-11.86854677	-10.64336	-11.8685468	-10.64335995
Number of Guests	0.03223051	0.002951408	10.92039	7.91316E-27	0.026441488	0.03801954	0.026441488	0.038019542
Quantity of Food	0.02655602	0.003166535	8.386462	1.08356E-16	0.020345036	0.03276701	0.020345036	0.032767009

In the second iteration of the model, *Number of Guests* was introduced as a new predictor variable, based on the logical assumption that a higher guest count would result in more food being prepared and, potentially, more food wastage. This addition improved the model's adjusted R^2 from 0.750 to 0.766, and slightly lower standard error of 5 to 4.88, indicating a stronger explanatory power. Although *Number of Guests* and *Quantity of Food* showed a high positive correlation ($r = 0.7769$), the value remains below the commonly cited multicollinearity threshold of 0.80, making the inclusion of both variables statistically justifiable (*Stataiml*, 2024). Their combined effect enhanced the model's predictive performance without introducing significant collinearity concerns. **(Refer to APPENDIX II: *Correlation matrix for predictor variables*)**

Subsequently, statistically insignificant predictors such as *Type of Food*, *Event Type*, and *Geographical Location* were removed to simplify the model and reduce the risk of overfitting. The final model remains strong while being more interpretable and generalizable. Notably, the confidence intervals for key predictors have become slightly narrower, suggesting greater precision and reliability in the parameter estimates. This refined model balances statistical robustness with practical applicability, making it more suitable for real-world food wastage prediction.

Hence, the final equation for predicting food wastage amount in kilograms:

Food Wastage Amount (kg) = 16.576 + (−1.009 × PreparationMethod_Buffet) + (−4.319 × PreparationMethod_FingerFood) + (−12.978 × Pricing_Low) + (−11.256 × Pricing_Moderate) + (0.032 × Number of Guests) + (0.0266 × Quantity of Food)

Limitations & Assumptions of Model

Limitations:	Assumptions:
● The dataset was sourced from a single Indian restaurant, which may limit the generalizability of the model to other settings, particularly Singapore's diverse F&B context. ● External variables such as event duration, menu complexity, or food spoilage rates were not captured, potentially omitting factors that also influence food waste. ● Cultural and operational differences across regions may affect how applicable the model is when applied beyond the original dataset. ● By removing statistically insignificant variables, the model may have excluded subtle influences that could contribute to improved accuracy in more complex settings.	● The model assumes linear relationships between the independent variables and food wastage. ● Dummy variable encoding assumes appropriate baseline categories for interpretation. ● The residuals are assumed to be normally distributed and exhibit constant variance (homoscedasticity). ● Each observation in the dataset is assumed to be independent of the others. ● The data used is assumed to be clean, free from major errors, and representative of typical F&B operations.

Conclusion

The final regression model offers a practical equation to predict food wastage based on statistically significant factors: *Preparation Method (Buffet, Finger Food, Sit Down Dinner)*, *Pricing (High, Moderate, Low)*, *Number of Guests*, and *Quantity of Food*. This equation shows that food waste increases with both the number of guests and the quantity of food prepared, which aligns with intuitive expectations. Conversely, lower pricing tiers and informal preparation methods (e.g., buffets and finger food) are linked to reduced wastage, likely due to simpler menus or more efficient portioning. To evaluate the model's performance, we relied on adjusted R^2 and standard error of regression, as these provide a more accurate reflection of model fit and predictive precision than R^2. With an adjusted R^2 of 0.766 and a standard error of ±4.88 kg, the model demonstrates strong predictive capability while remaining interpretable and generalisable.

Peer Contribution

Name	Student ID	DTP Project Contribution
Davin Handreas Chen	1009595	Analyze numerical data. Build and improve the MLR model. Write and finalize the MU report.
Thong Zi Qi	1009160	Build MLR model in DDW, evaluate and improve model. Refactor the Excel sheets in MU.
Catherine Laura Danandjaja Antoni	1009263	Data Cleaning for DDW. Interpret the DDW model. Refactor the MU report.
Ryan Leow Heng Kuan	1009273	Source for numerical data in MU. Background check and write the MU report. Refactor the MU report.
Wong Jun Han Brayden	1009259	Clean the raw data in MU, and analyse categorical data. Build the interactive model in Excel.

APPENDIX I: Dummy Variables & Baselines

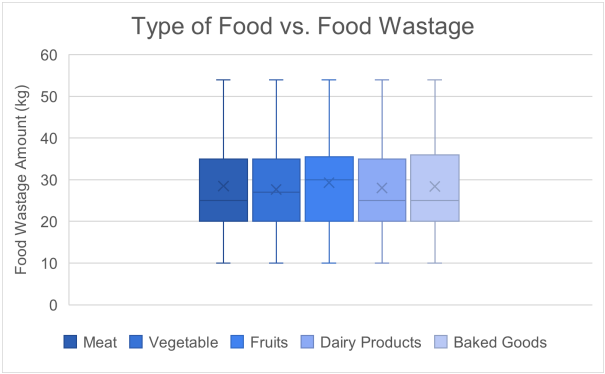
Categorical Variable	**Baseline**	**Dummy Variable Used**
Type of Food	Baked goods	Meat, Fruits, Dairy Products, Vegetables
Pricing	High	Low, Moderate
Event Types	Birthday	Corporate, Wedding, Social Gathering
Geographical Location	Rural	Urban, Suburban
Preparation Method	Sit-down Diner	Buffet, Finger Food

APPENDIX II: Correlation matrix for predictor variables

	ToF_Meat	ToF_Vegetables	ToF_Fruits	ToF_DairyProducts	ET_Corporate	ET_Wedding	ET_SocialGathering	PM_Buffet	PM_FingerFood	P_Low	P_Moderate	GL_Suburban	GL_Urban	Number of Guests	Quantity of Food
ToF_Meat	1														
ToF_Vegetables	-0.217106	1													
ToF_Fruits	-0.274092	-0.182278201	1												
ToF_DairyProducts	-0.299057	-0.198880285	-0.251082	1											
ET_Corporate	0.036065	0.021637453	-0.012995	0.011771677	1										
ET_Wedding	0.045149	0.01307557	0.019929	-0.041429914	-0.36525169	1									
ET_SocialGathering	-0.082905	-0.027205154	-0.013242	0.037198989	-0.37189211	-0.341365	1								
PM_Buffet	0.024299	0.004047227	0.001292	-0.005157468	0.003165578	-0.1035719	0.068715837	1							
PM_FingerFood	-0.035893	-0.02297581	0.004552	0.018987743	0.048386105	-0.1448972	0.033689392	-0.355397	1						
P_Low	0.023485	0.01405582	-0.031818	0.006364742	-0.00539226	0.01004499	0.000818566	0.0988759	-0.00726266	1					
P_Moderate	-0.008022	0.003122026	-0.014836	-0.000513968	0.026664669	-0.0523775	0.018465765	-0.104483	0.241008762	-0.4485	1				
GL_Suburban	-0.000318	-0.025780288	0.005811	3.01707E-05	-0.05739981	-0.0971879	0.034488875	-0.044333	-0.078841212	-0.0703	0.05526552	1			
GL_Urban	0.019324	-0.005912381	-0.012304	0.01986806	0.193914312	0.048317	-0.124500311	0.0211662	-0.118767488	0.10658	-0.08095225	-0.632685227	1		
Number of Guests	-0.005262	-0.031935272	0.03782	-0.013867771	-0.03599502	0.0315166	-0.035458651	0.0371995	-0.262029124	-0.1167	-0.26772761	-0.002950095	0.064266	1	
Quantity of Food	0.020583	-0.033845484	0.037518	-0.025350699	-0.01302081	0.04817636	-0.037107807	0.0397981	-0.381227063	-0.0628	-0.31335906	0.024495924	0.088537	0.776941263	1

APPENDIX III: Box Plot Analysis of Categorical Datas

Chart 1. Box Plot: Type of Food vs. Food Wastage



All food categories (*Meat, Vegetables, Fruits, Dairy Products, Baked Goods*) display similar median wastage levels, indicating no major differences across types. *Baked Goods* and *Fruits* exhibit slightly wider spreads, suggesting more variability, but there are no extreme outliers or skewness.

Chart 2. Box Plot: Pricing vs. Food Wastage

Pricing shows a clear pattern: higher-priced events have greater median wastage and a wider spread. The presence of a few mild outliers in the high-pricing category likely reflects realistic scenarios such as over-catering. These outliers fall within reasonable bounds and do not distort the distribution. This variable is expected to be influential in the regression.

Chart 3. Box Plot: Type of Event vs. Food Wastage

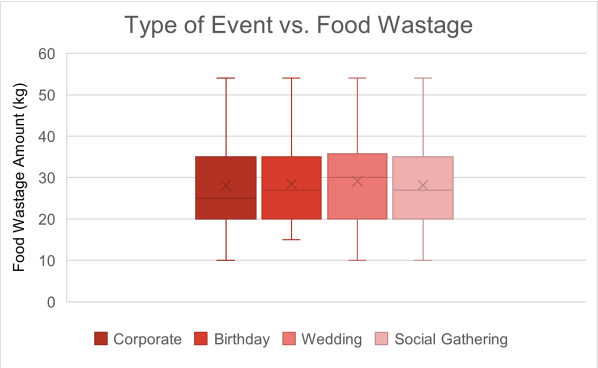
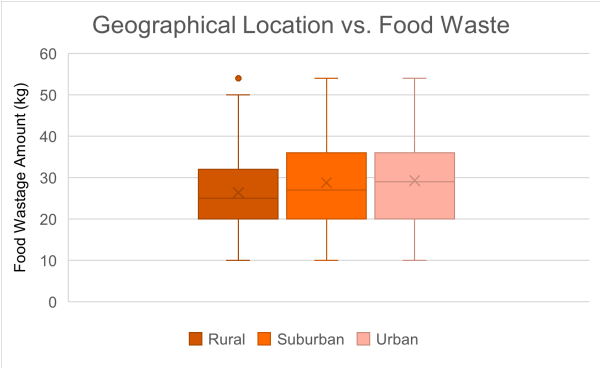


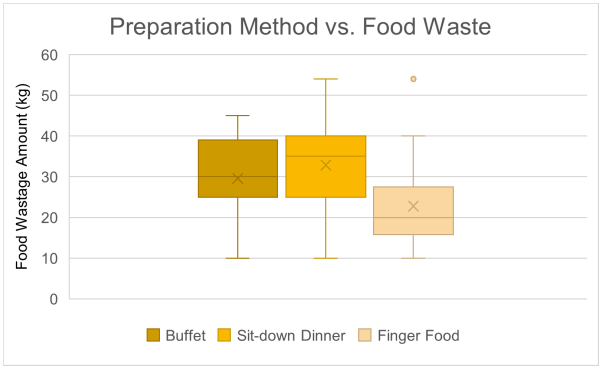
Chart 4. Box Plot: Location vs. Food Wastage



Corporate, Birthday, Wedding, and *Social Gathering* events show largely consistent median food wastage. The *Wedding* category presents a slightly broader interquartile range, possibly due to event size differences, but overall distribution remains balanced. No significant outliers were observed, meaning event type can be retained as-is for regression analysis.

Rural, Suburban, and *Urban* locations have comparable median wastage, though Urban events show a slightly larger spread. A mild outlier is seen in the *Rural* group, but it does not distort the overall pattern. The distribution is fairly balanced, and geographical location is suitable for inclusion in the multilinear regression model.

Chart 5. Box Plot: Prep. Method vs. Food Wastage



Buffet and *Sit-down Dinner* show similar medians, though *Buffet* has a wider range. *Finger Food* also displays moderate variability. One mild outlier each appears in the *Sit-down Dinner* and *Finger Food* categories. These are considered realistic deviations due to practical event factors and should be kept. The preparation method appears meaningful and can be used in the model.

APPENDIX IV: Data Analysis for Number of Guests vs. Food Wastage Regression

Regression Statistics	
Multiple R	0.6187
R Square	0.3828
Adjusted R Square	0.3824
Standard Error	7.9400
Observations	1609.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.0000	62830.6479	62830.6479	996.6221	0.0000
Residual	1607.0000	101311.0724	63.0436		
Total	1608.0000	164141.7203			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.2068	0.9596	-1.2576	0.2087	-3.0890	0.6754	-3.0890	0.6754
Number of Guests	0.0934	0.0030	31.5693	0.0000	0.0876	0.0992	0.0876	0.0992

APPENDIX V: Data Analysis for Quantity of Food vs. Food Wastage Regression

Regression Statistics	
Multiple R	0.6172
R Square	0.3810
Adjusted R Square	0.3806
Standard Error	7.9515
Observations	1609.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.0000	62537.3756	62537.4	989.1069412	1.36E-169
Residual	1607	101604.3447	63.2261		
Total	1608	164141.7203			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-11.3277	1.2798	-8.8511	2.23678E-18	-13.8379	-8.8174	-13.8379	-8.8174
Quantity of Food	0.0966	0.0031	31.4501	1.36E-169	0.0906	0.1026	0.0906	0.1026

APPENDIX VI: 1st iteration of multilinear regression

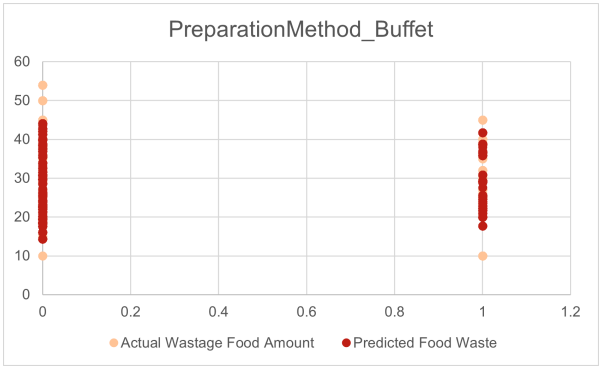
Regression Statistics	
Multiple R	0.867279419
R Square	0.752173591
Adjusted R Square	0.749996948
Standard Error	5.051718284
Observations	1609

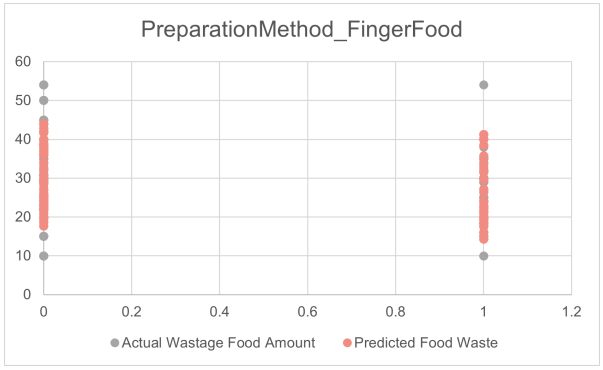
ANOVA

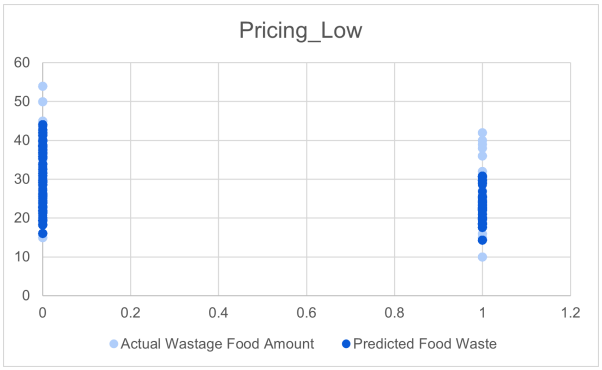
	df	*SS*	*MS*	*F*	*Significance F*
Regression	14	123463.0673	8818.79	345.5658199	0
Residual	1594	40678.65304	25.5199		
Total	1608	164141.7203			

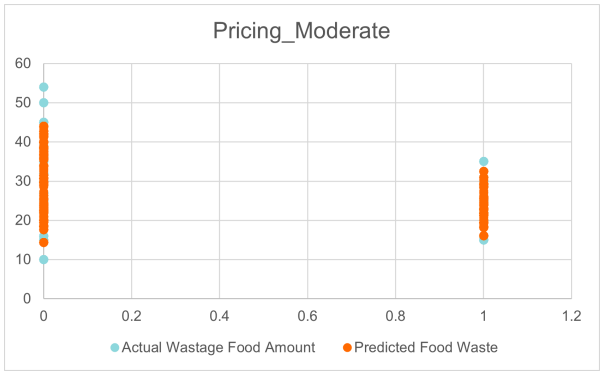
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	16.85105479	1.145061461	14.7163	4.61635E-46	14.6050702	19.0970394	14.60507015	19.09703942
ToF_Meat	-0.099419752	0.369736003	-0.26889	0.788046137	-0.82463967	0.62580017	-0.82463967	0.625800169
ToF_Vegetables	-0.408786187	0.444873751	-0.91888	0.358296665	-1.28138529	0.46381292	-1.28138529	0.463812922
ToF_Fruits	-0.061600198	0.394908197	-0.15599	0.876063711	-0.8361942	0.71299381	-0.8361942	0.712993806
ToF_DairyProducts	-0.275565757	0.380114711	-0.72495	0.46858653	-1.02114303	0.47001151	-1.02114303	0.470011515
ET_Corporate	-0.342651155	0.372209177	-0.92059	0.357405073	-1.07272209	0.38741978	-1.07272209	0.38741978
ET_Wedding	-0.803965614	0.38708149	-2.07699	0.037962024	-1.5632079	-0.0447233	-1.5632079	-0.04472333
ET_SocialGathering	0.099430596	0.374323509	0.26563	0.790560498	-0.6347875	0.8336487	-0.6347875	0.833648695
PM_Buffet	-0.922220654	0.373639273	-2.46821	0.013683316	-1.65509666	-0.1893447	-1.65509666	-0.18934465
PM_FingerFood	-4.023192485	0.320380494	-12.5575	1.47272E-34	-4.65160388	-3.3947811	-4.65160388	-3.39478109
P_Low	-13.68408887	0.337073908	-40.5967	5.0427E-248	-14.3452436	-13.022934	-14.3452436	-13.0229341
P_Moderate	-11.8469382	0.32008699	-37.0116	6.0327E-217	-12.4747739	-11.219102	-12.4747739	-11.2191025
GL_Suburban	0.494506233	0.341803569	1.44676	0.148161993	-0.17592552	1.16493799	-0.17592552	1.164937987
GL_Urban	0.978606202	0.372297139	2.62856	0.00865704	0.24836273	1.70884967	0.248362733	1.708849671
Quantity of Food	0.051104898	0.002251105	22.7021	4.3294E-99	0.04668946	0.05552034	0.046689461	0.055520335

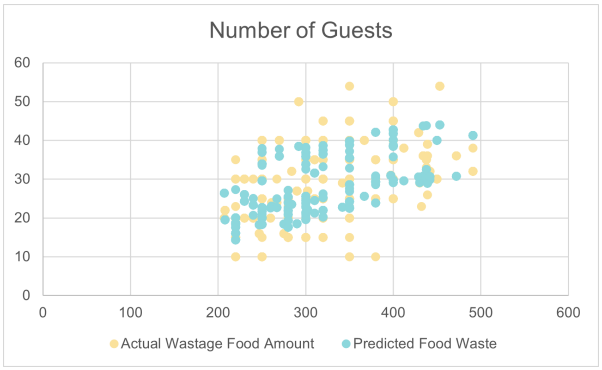
APPENDIX VII: Comparison of Predicted and Actual Food Wastage

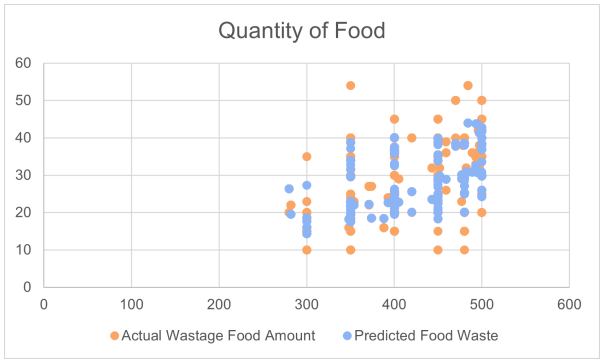












Resources

1. Council, S. E. (2019). Annual Report 2019 - Heart Reset, Green Pivot. https://sec.org.sg/pdf/annual-reports/annual-report-2019.pdf
2. Hannibal, T. (2023, May 28). *Food wastage data in Restaurant*. Kaggle. https://www.kaggle.com/datasets/trevinhannibal/food-wastage-data-in-restaurant
3. Loy. (2015, January 18). *Dummy Variables*. RPubs. https://www.rpubs.com/Loy/dummyvariables
4. Stataiml. (2024, July 2). Thresholds for detecting multicollinearity. https://stataiml.com/posts/60_multicollinearity_threshold_ml/