Table of Contents

1.Business Background

1.1 Overview of Choco Meridian:

Founded in 1998 in London, Choco Meridian crafts exquisite chocolates inspired by global flavors, blending tradition and innovation for a unique experience. Currently we are operating in Mexico, Singapore, Spain, UK, and US.

Figure 1.1: Geographical Mapping

Choco Merian is currently operating in Mexico, Singapore, US, Spain, and UK.

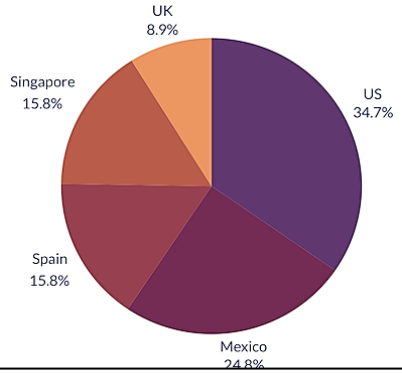


Figure 1.2 Regional Growth of Choco Meridian

As of 2014, Choco Meridian, had the highest market share in US, followed by Mexico, Singapore, and Spain with the shares of 34.6%, 24.5%,15.7%,15.7% respectively leaving UK had the lowest (9.4%)

1.2 Market Research

1.2.1 Current Global Market Share

According to ICCO (2015) Western Europe leads the global chocolate market at 35%, followed by North America (21%) and Asia Pacific (14%). Statista (2024)

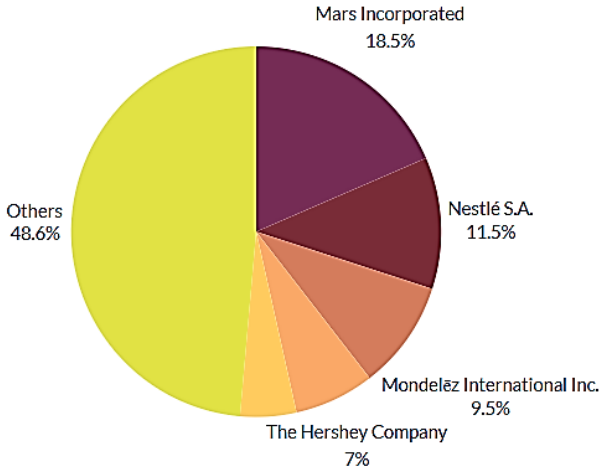


Figure 1.3: Current Market Share

1.2.2 Competitor Analysis

In 2014, the global chocolate industry was dominated by several key players. The top five key competitors from the current market are:

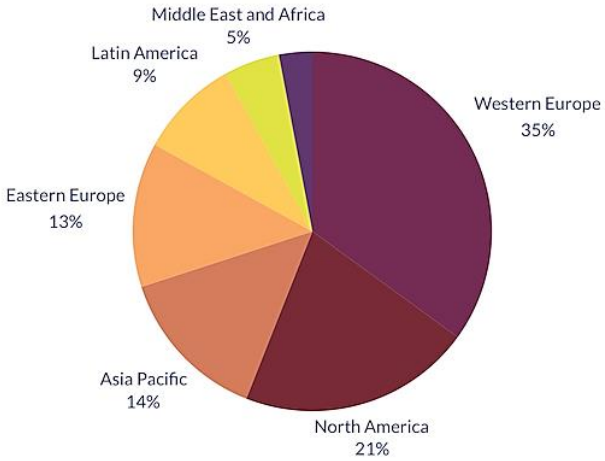


Figure 1.4: Current Market Share

- Mars Incorporated
- Nestlé S.A.
- Mondelēz International Inc.
- The Hershey Company
- Ferrero International S.A.

Western Europe leads the global chocolate market with 35%, followed by North America (21%), Asia Pacific (14%), Eastern Europe (13%), and Latin America (9%).

2. Data and Methodology

1: DATA PREPROCESSING

- **Count Analysis**

The dataset includes 60 data points evenly distributed across all five locations for balanced analysis.

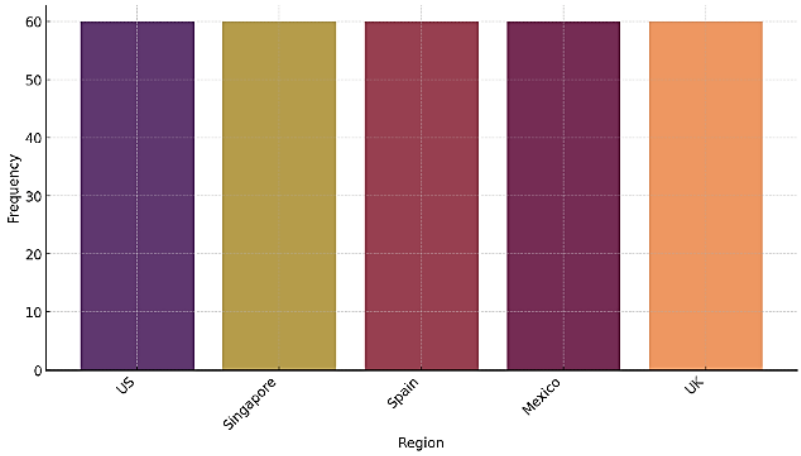


Figure 1.5: Frequency of Data by Region

- **Renaming of columns:** Below are the initial and renamed column names.

Original Column	Renamed Column	Description
Sales (units)	Sales	Number of units sold.
Price (Â£.k)	Price	Cost of the product sold.
Ad1 (GRP)	Ad1	TV advertisement impact score.
Ad2 (No_of_banners)	Ad2	Banner advertisement impact score.
Prom (No_of_stores)	No_of_stores	Total stores selling product.
Wage (Perc.)	Wage	Income group in the region.
Time	Time	Timeline of the sales period.
Product	Product	Specific product being analyzed.
Region	Region	Geographic market of sales activity.
Month	Month	Month of sales transaction.
Year	Year	Year of sales transaction data.

2: EDA (Exploratory Data Analysis)

Performing univariate analysis, i.e, histogram plot for all features and checking correlation matrix revealed anomalies and skewness which directed us to perform feature engineering.

Figure 2.2: Histogram Chart

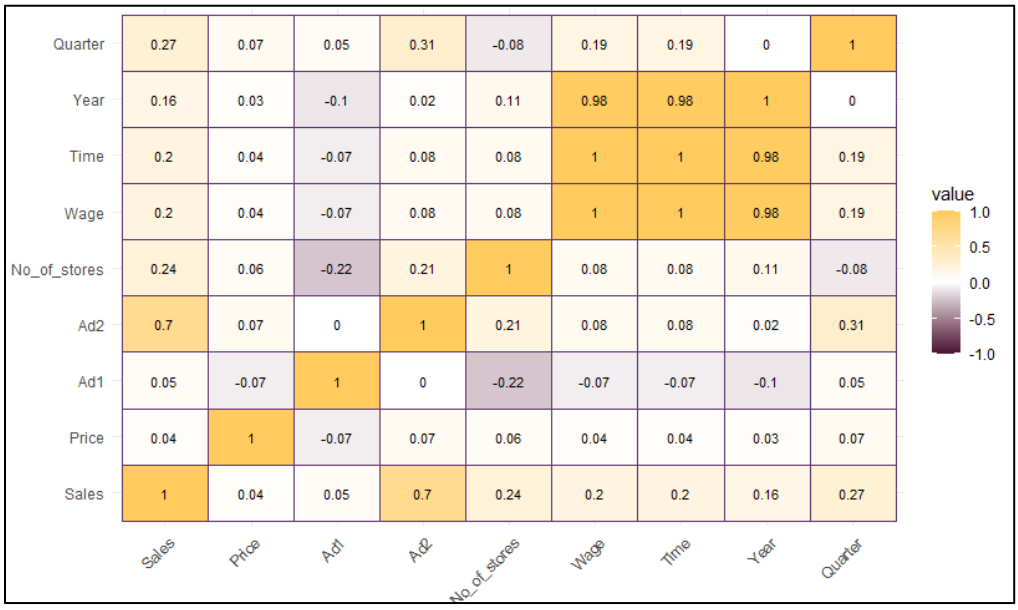


Figure 2.2: Correlation Matrix Before Feature Engineering

3: FEATURE ENGINEERING

We produced the following interaction terms and logarithmic transformations to improve the forecast accuracy.

Features Added	Description
Ad1_sq	To address linearity.
Ad2_sq	To address linearity.
Ad1_Ad2	To capture interacted effect.
Ad2_Prom	To capture interacted effect.
Log_Ad1	To address high skewness and 0 values.
Log_Ad2	To address high skewness.
Log_Prom	To address high skewness.
Log_Sales	To address high skewness and outliers.
Ad1_zero_indicator	Created binary indicator for no advertising (zero value).

Checked the correlation for all the features and obtained the following results.

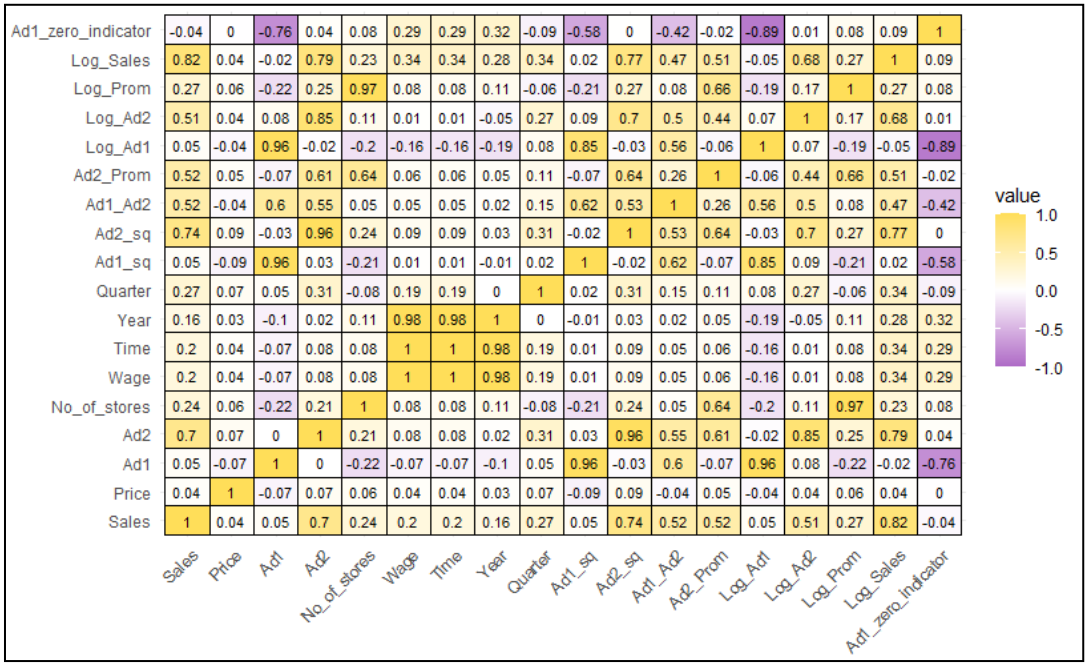


Figure 2.3: Correlation Matrix After Feature Engineering

4: PCA (Principal Component Analysis)

PCA reduced dimensionality by transforming correlated variables into uncorrelated principal components, preserving maximum variability (80%).

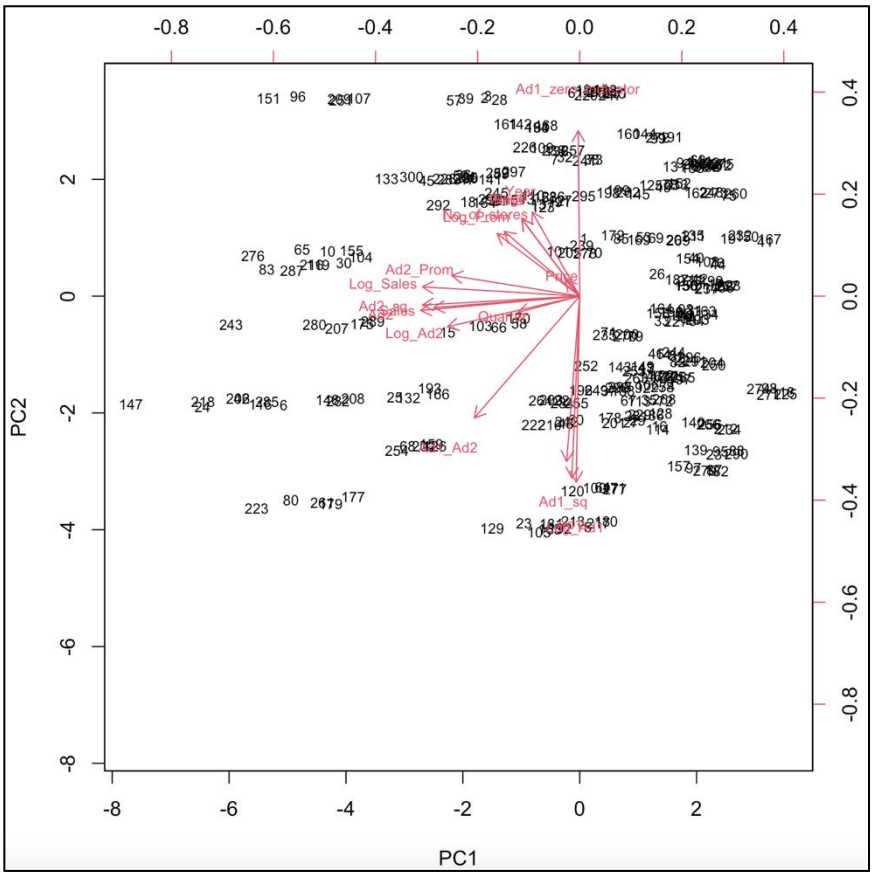


Figure 2.4: Plot for PCA Model

Biplot highlights key predictors, relationships and outliers for insights

5: TESTING AND HANDLING OUTLIERS

To handle the outliers, we used boxplots as they visually display the key aspects of a dataset:

- Central tendency.
- Spread.
- Extreme values (Outliers)

Analyzing this, we scaled some features to handle the outliers and avoid biased analysis.

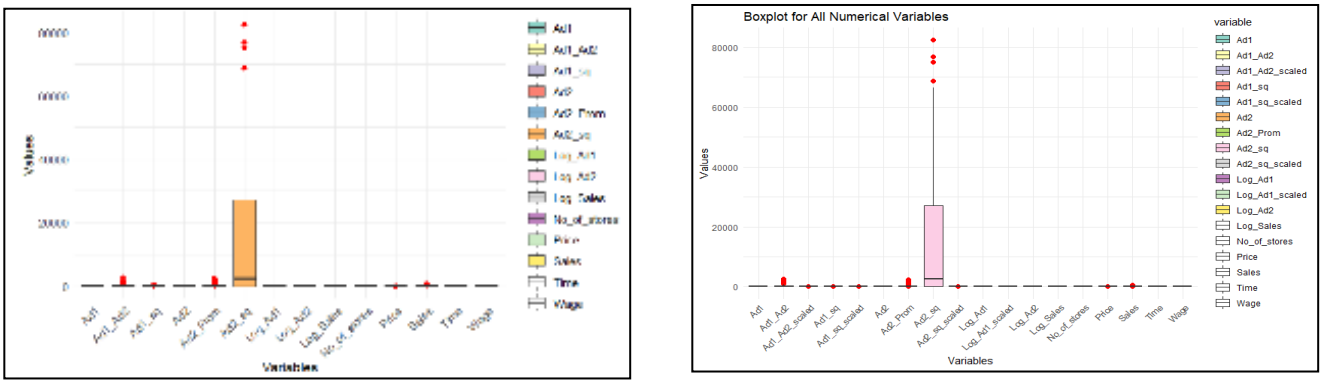


Figure 2.5: Boxplot Before Scaling and After Scaling

6: SPLITTING THE DATA

The dataset was initially split regionally into training (80%) and testing (20%).

7: MODEL PREPARATION AND VISUALISATION

Various models tested across all the regions in the following order :

Linear Regression Model ➡ Time Series ➡ ARIMA ➡ PCA ➡ Multiple Regression Model.

E.g. of error metrics evaluated for model's performance for the US.

Model	Multiple R^2	Adjusted R^2	Train RMSE	Test RMSE	MAPE	MAE
Linear Regression Model	0.0148	-0.0066	149.42	101.09	149.76%	76.67
Time Series Model	-	-	108.81	-150.73	72.29%	109.88
ARIMA Model	-	-	133.18	-	175.39%	98.58
PCA Model	0.9512	0.9467	0.21	0.22	20.55%	9.30
Multiple Regression Model	0.9065	0.8954	0.30	0.30	23.31%	33.47

Similarly, we evaluated the models for the remaining regions as well.

8: MODEL EVALUATION

We predicted sales for testing datasets and evaluated the actual and predicted sales, based on error matrix.

9: FINAL ANALYSIS

Model coefficients were analyzed to determine each predictor's impact on sales, with exp(predictions) - 1 being used to convert predictions back to the original scale.

3. Managerial Summary

3.1 How many units sold are associated with each of the three marketing activities?

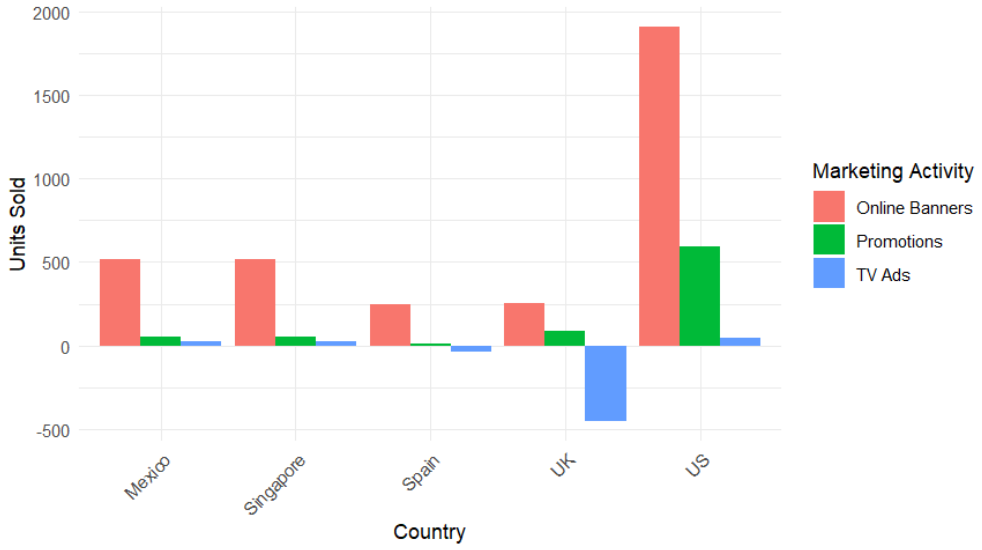


Figure 3.1: Sales Contribution by Marketing Activity Across Regions

Region	TV Advertisements	Online Banner Advertisements	Promotions
Mexico	167.43	1513.24	238.52
Spain	-38.41	633.36	30.48
US	45.56	2333.74	351.01
UK	-1126.21	489.87	39.12
Singapore	134.6	927.66	66.56

Graph 3.1 depicts that promotions and online banners dominate sales across countries.

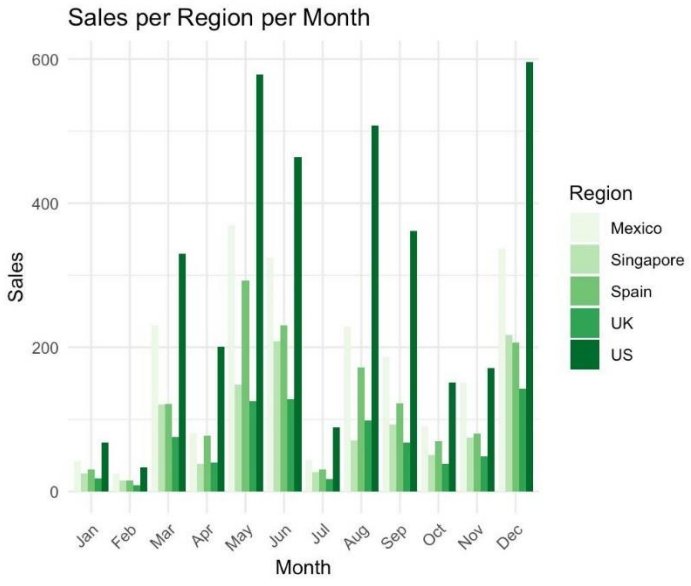
Online banners have proven to be the most effective activity, particularly in the US. This data suggests focusing on enhancing promotions and online banner strategies for better reach and impact across all regions.

3.2 Our TV ads cost us £2,000,000 a year and our Banners £500,000 a year (in total for all regions). Which one is more cost-effective?

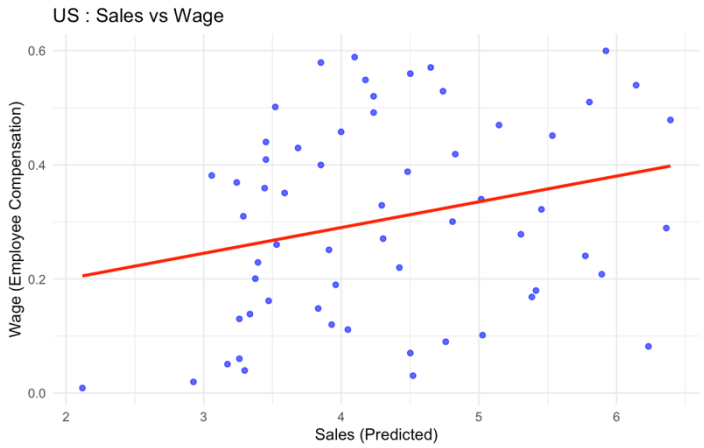
Advertising Type	Yearly budget (£)	Total impact score	Cost per unit Impact (£)	Effectiveness
TV Ads	2,000,000	-817.13	2447.60	Highly costly, poor value
Banner Ads	500,000	5897.97	84.78	Excellent value, significant impact

Budget Allocation Insight: TV ads account for 80% of the advertising budget but contribute a negligible positive impact. Allocating more budget to banner ads would deliver far better results.

3.3 Our sales often show a lot of variation. Can you explain to us possible sources of the variation, other than the marketing activities?



The chart shows seasonal sales trends, with peaks in May and December due to holidays like Mother's Day and Christmas, and dips in January-February from post-holiday spending fatigue.



The data highlights a positive link between regional wages and sales, with higher-wage regions like the UK and US outperforming lower-wage regions.

3.4. Could you provide a prediction of the sales of the next month for all countries/regions?

REGION	ANALYSIS
US	63.09 units → Strong demand and growth.
Mexico	37.74 units → Steady potential.
Spain	22.33 units → Moderate sales.
Singapore	16.48 units→ Consistent engagement
UK	13.83 units→ Potential challenges

These predictions were generated using a multiple regression model.

4. Technical Summery

4. 1 Question 1: How many units sold are associated with each of the three marketing activities?

REGION	TV ADS	ONLINE BANER ADS	PROMOTIONS
Mexico	167.33	1513.24	238.52
Spain	-38.41	633.36	30.48
US	45.56	2333.74	351.01
UK	-1126.21	489.87	39.12
Singapore	134.6	927.66	66.56

- We split the data for each region and performed stepwise selection.
- Built a regression model for each region with best variable selection based on the results we received also performed all the model assumptions i.e. Linearity, Homoscedasticity, Normality, Q-Q Plot for Normality of Residuals and Multicollinearity.

Multiple Regression Model Based On Stepwise Selection On Full Model	Country
lm (Log_Sales ~ Log_Ad1_scaled + Ad2 + Wage + Log_Prom + Quarter , data = train_data)	Mexico
lm (Log_Sales ~ Ad1 + Ad2 + Log_Prom + Quarter + Ad1_Ad2_scaled, data = train_data)	Spain
lm (Log_Sales ~ Ad1 + Ad2 + Wage + Log_Prom + Quarter , data = train_data)	US
lm (Log_Sales ~ Ad2 + Wage + Log_Prom + Quarter + Ad1_Ad2_scaled + Time + Ad1_sq_scaled, data = train_data)	UK
lm (Log_Sales ~ Log_Ad1_scaled + Ad2 + Wage + Log_Prom + Quarter , data = train_data)	Singapore

- The below figure shows the Error matrix from each regression model to get the very low mean absolute percentage error with excellent multiple and adjusted R squared value.

Country	Multiple R-squared	Adjusted R-squared	Test RMSE	Train RMSE	MAPE
Mexico	0.9351	0.9273	0.2361968	0.2498748	21.43016 %
Spain	0.8867	0.8732	0.2988392	0.33043	26.52%
US	0.902	0.8903	0.2778966	0.3099888	24.25%
UK	0.902	0.8903	0.2778966	0.3099888	24.25%
Singapore	0.9568	0.9517	0.2470548	0.2305768	22.89%

- Calculation of the units sold for Each region was based on the below formula
- Extract coefficients (beta_log_ad1, beta_ad2, beta_log_prom) for the log-scaled or direct predictors of advertising and promotions.
- Calculate the mean values of predictors (e.g., Ad1, Ad2, No_of_stores) from the training dataset.

```
ImpactAd1 = (exp (log_ad1 - log (Mean_Ad1 + 1)) - 1) * Mean_Ad1
ImpactAd2 = exp(βad2 * Mean_Ad2) - 1
ImpactProm = (exp (βlog_prom - log (Mean_Prom + 1)) - 1) * Mean_Prom
```

Predicting Units Sold for the testing dataset

Use the calculated impacts and predicted sales (test_data$predicted_sales) to estimate the number of units sold attributed to each activity:

Marketing Activity	Formula	Description
TV Advertisements (Ad1)	Units_Sold_Ad1 = Σ (predicted_sales · Impact_Ad1)	Total units sold attributed to TV advertisements, based on predicted sales and calculated impact.
Online Banners (Ad2)	Units_Sold_Ad2 = Σ (predicted_sales · Impact_Ad2)	Total units sold attributed to online banner advertisements.
Promotions (Prom)	Units_Sold_Prom = Σ (predicted_sales · Impact_prom)	Total units sold attributed to promotions in stores.

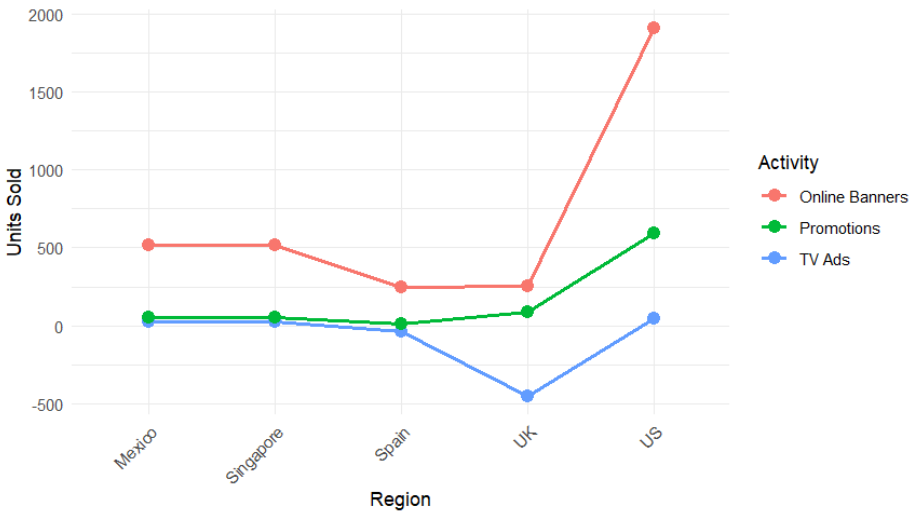


Figure 4.1: Impact of Marketing Channels in Sales

This framework converts the model's coefficients into actionable estimates, showing the impact of each marketing channel on sales. By summing up contributions across the test dataset, it provides a clear measure of each activity's effectiveness in driving sales.

4.2 Question 2: Our TV ads cost us £2,000,000 a year and our Banners £500,000 a year (in total for all regions). Which one is more cost-effective?

Region = c("Mexico", "Spain", "US", "UK", "Singapore")
TV_Ads = c(mexico_ad1, spain_ad1, us_ad1, uk_ad1, sing_ad1)
Online_Banners = c(mexico_ad2, spain_ad2, us_ad2, uk_ad2, sing_ad2)
Promotions = c(mexico_prom, spain_prom, us_prom, uk_prom, sing_prom

- These numbers measure how impactful each ad type is in driving sales across multiple regions. Positive values contribute to sales, and negative values indicate adverse effects.
- Calculated the sum of ad1 i.e. TV ads and ad2 i.e. Banners for each region .

Total Effectiveness of TV Ads (Ad1) = ∑ (TV ads effectiveness values)

Total Effectiveness of Banner Ads (Ad2) = ∑ (Banner ads effectiveness values)

- The cost per unit effectiveness is calculated to evaluate how much money is spent for each unit of effectiveness.

$$\text{Cost per Unit Effectiveness of TV Ads (Ad1)} = \frac{\text{Total Cost}}{\text{Total Effectiveness of TV Ads}}$$

$$\text{Cost per Unit Effectiveness of Banner Ads (Ad2)} = \frac{\text{Total Cost}}{\text{Total Effectiveness of Banner Ads}}$$

Advertisement	Cost
Total_TV_Effect	-817.13
Total_Banner_Effect	5897.97
Cost_Per_Effect_TV	-2,447.60
Cost_Per_Effect_Banner	84.78

The table above shows that Banner ads are significantly more cost-effective because the cost per unit effectiveness is much lower compared to TV ads.

4.3 Question 3: Our sales often show a lot of variation. Can you explain to us possible sources of the variation, other than the marketing activities?

Based on the regression model we concluded that Wage and Quarter significantly affect sales in region.

Region	Regression Model
Mexico	Log_Sales=β0 +β1* Log_Ad1_scaled+β2 *Ad2+β3 *Wage+β4* Log_Prom+β5 *Quarter+ϵ
US	Log_Sales=β0 +β1 *Ad1+β2* Ad2+β3* Wage+β4* Log_Prom+β5* Quarter+ϵ
Spain	Log_Sales=β0 +β1* Ad1_sq_scaled+β2 *Ad2+β3* Log_Prom+β4* Quarter+ϵ
UK	Log_Sales=β0 +β1* Ad2+β2 *Wage+β3* Log_Prom+β4* Quarter+β5 *Ad1_Ad2_scaled+β6 *Time+β7* Ad1_sq_scaled+ϵ
Singapore	Log_Sales =β0 +β1* Log_Ad1_scaled+β2 *Ad2+β3* Wage+β4 *Log_Prom+β5* Quarter+ϵ

Most significant effect of wage on sales is provided by Singapore 3.204.

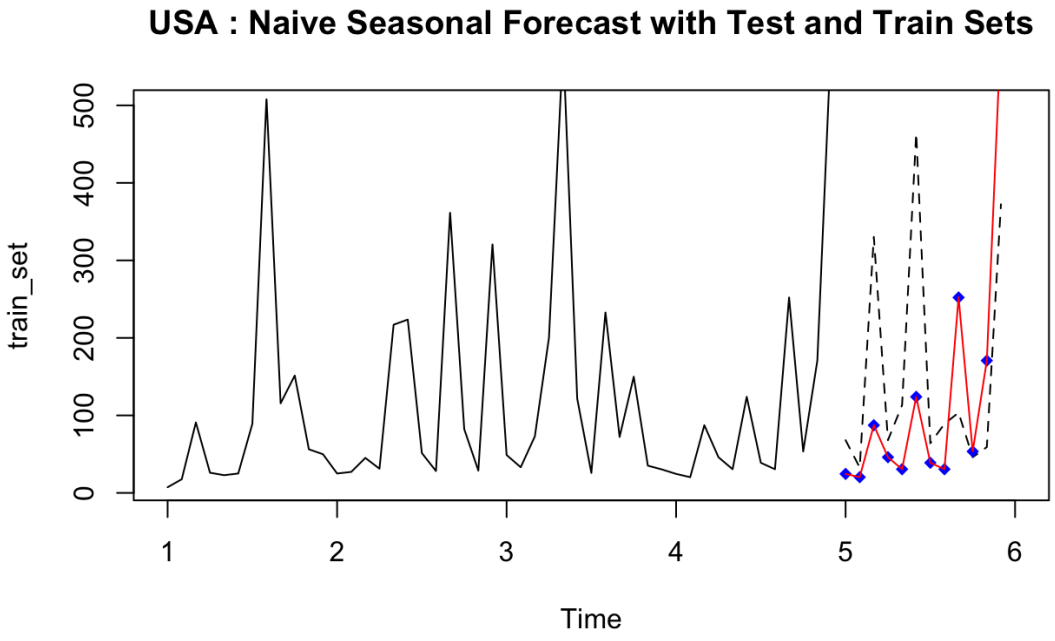
Region	Coefficient Of Wage	Interpretation
Mexico	1.558	Significant positive effect.
US	0.940	Moderate positive effect.
Spain	Not Included	Wage not used in the model.
UK	1.595	Exceptionally strong, but high variability.
Singapore	3.204	Very strong positive effect.

Quarter captures seasonal effects, with Spain showing the highest seasonal influence

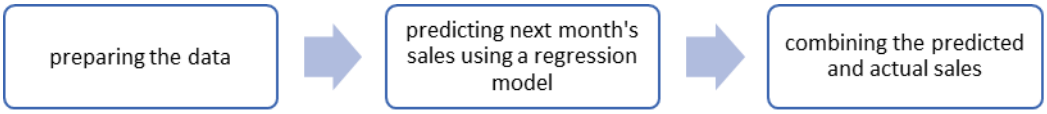
Region	Coefficient Of Quarter	Interpretation
Mexico	0.077	Marginally significant, minor effect (≈7.8% increase per unit).
US	0.093	Positive but borderline significant.
Spain	0.089	Statistically significant, indicating seasonal impact.
UK	0.083	Positive, but only marginally significant.
Singapore	0.077	Not statistically significant.

We analyzed time series of sales data and decomposed it using STL decomposition to get the trend, seasonal and remainder components for the time period and applied a naive seasonal forecasting method to understand if sales follow a repeating seasonal pattern across all region.

$$\mathbf{Y = Trend + Seasonal + Irregular}$$



4.4 Question 4: Could you provide a prediction of the sales of the next month for all countries/regions?



We split the data for all 5 regions i.e., UK, SPAIN, MEXICO, SINGAPORE, USA. And created the Regression model as explained in question 1.

1. DATA PREPARATION

We summarized the dataset to compute the mean values of numerical predictors. These means represent the average state of the dataset for the next month's prediction. For formula for column X is given below:

$$\text{Mean}(X) = \frac{\sum_{i=1}^{n} X_i}{n}$$

2. SCALING PREDICTORS

- Scaling ensures that predictors are normalized for the regression model.
- Predictors in our model are: Ad1_sq , Ad2_sq , Log_Ad1, Ad1_Ad2 where mean(X) is the mean of the predictor and std(X) is the standard deviation.

$$X_{\text{scaled}} = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

3. ADDING NEXT MONTHS' TIME VARIABLES

Temporal variables are added to capture the context of the next month i.e.:

Month: The specific month for the prediction = January
Year: The year of the prediction = 2015
Quarter: The quarter of the year = (e.g., Q1).
Season: A qualitative label = Winter

4. PREDICTING LOG SALES

$$\text{Log\_Sales\_Predicted} = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot X_i$$

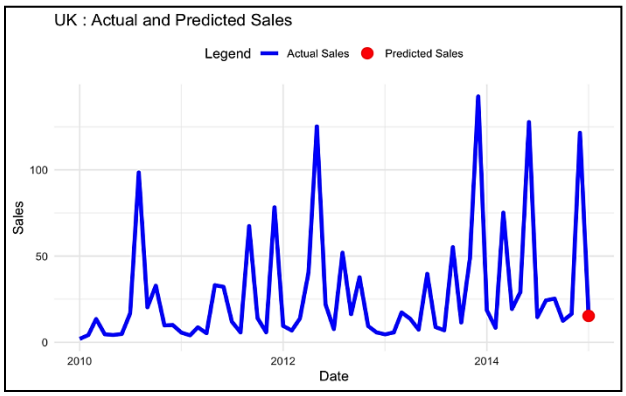
The regression model predicts sales in logarithmic form:
Where: β0 is the intercept and βi are coefficients of predictors Xi (e.g., Log_Ad1_scaled, Ad1_sq_scaled).

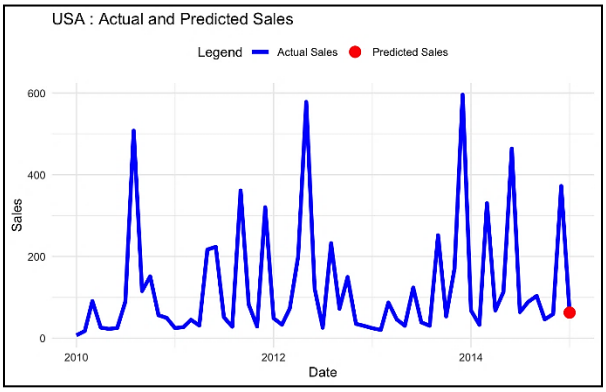
5. REVERSE LOG TRANSFORMATION

The logarithmic transformation is reversed to obtain actual predicted sales: And we have performed the above steps for each region to Predict the sales of 2015.

Country	Sales (Unit)
Mexico	37.74
Singapore	16.48
UK	13.83
Spain	22.33
US	63.09

6. VISUALISATION OF PREDICTED SALES IN 2015





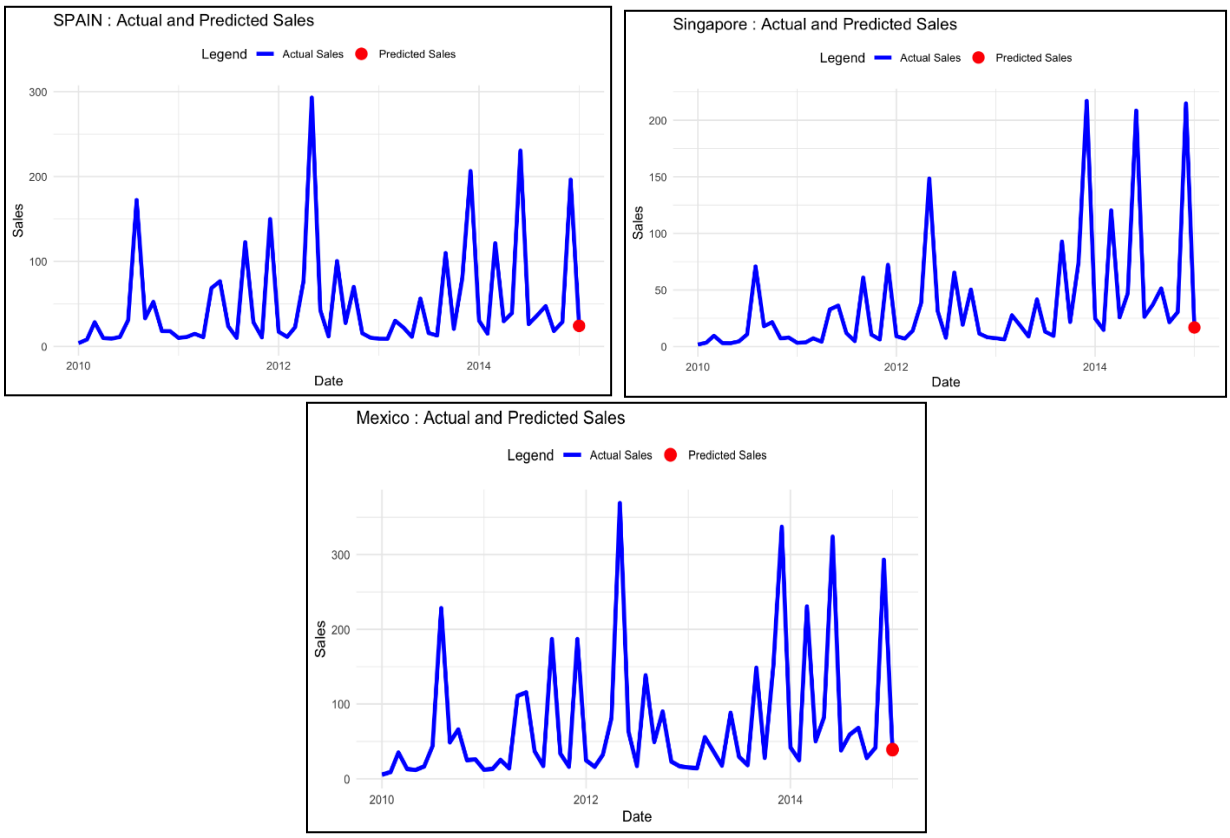


Figure 4.2: Forecasting for January by Region

5. Validation and Robustness Checks

5.1 Assumptions

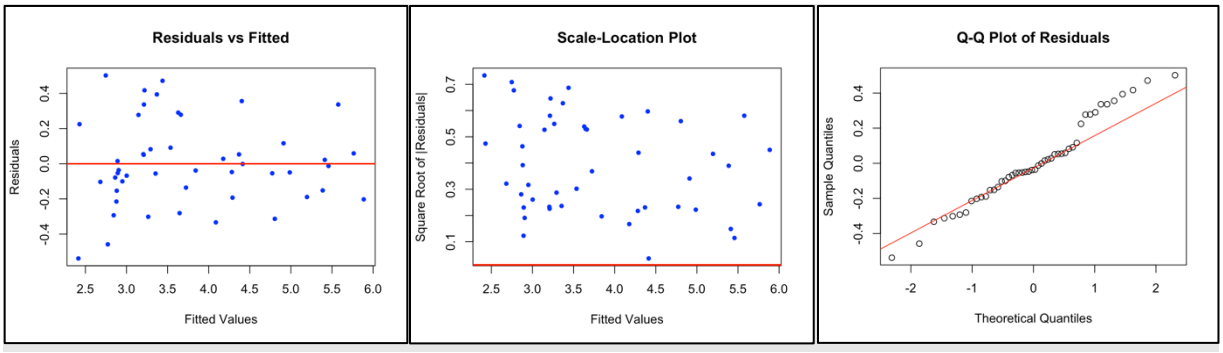
An optimal regression model should meet the following assumptions:

- Linearity
- Homoscedasticity
- Normality
- Multicollinearity
- Independence of residuals

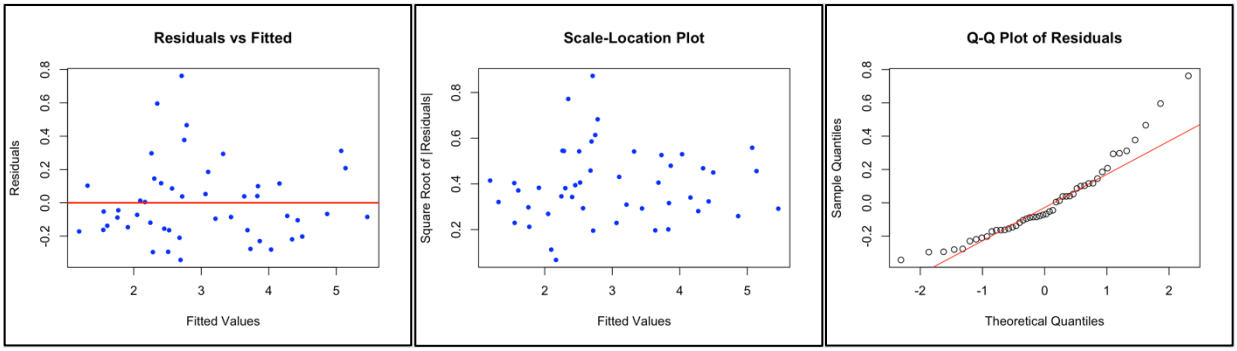
Since the multiple regression model satisfied majority of the above mentioned assumptions, it proved to be the most optimal across all the regions.

Below are the assumption plots for all the regions:

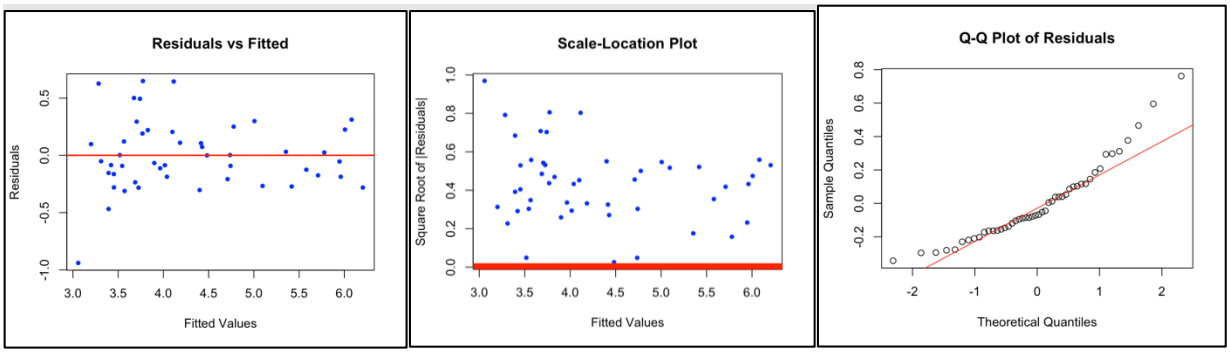
MEXICO



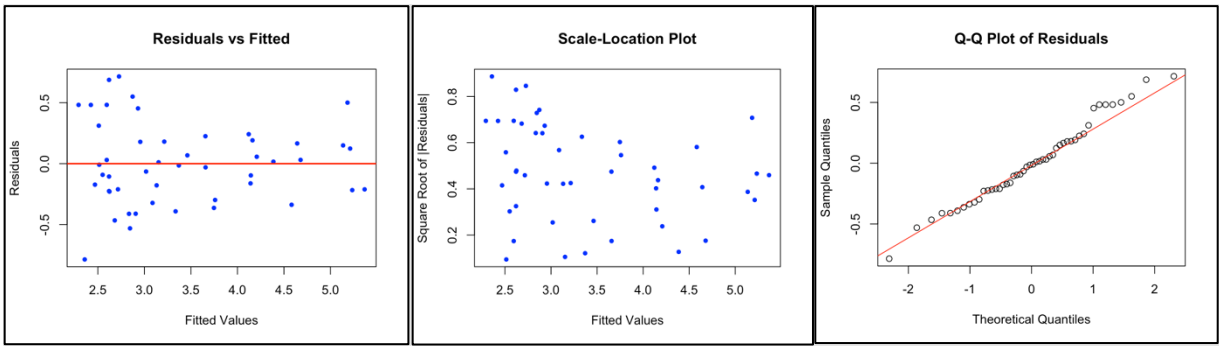
SINGAPORE



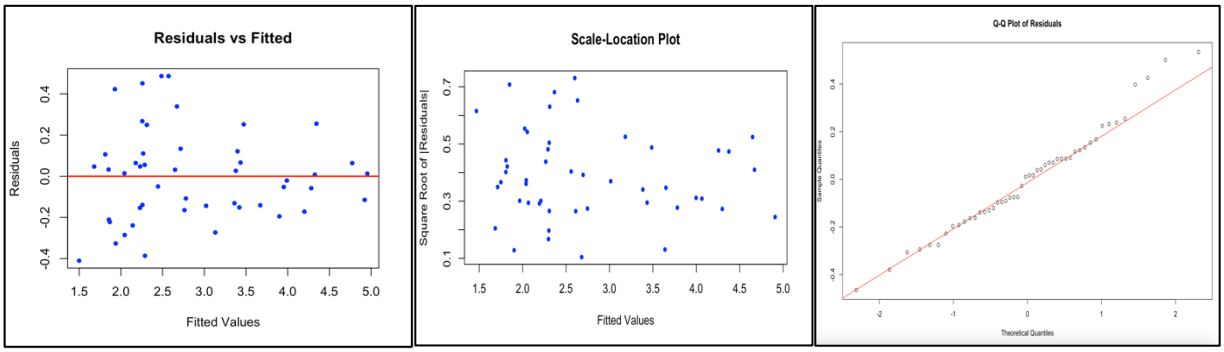
US



SPAIN



UK



After analyzing the above plots, it is clearly visible that all the models satisfied the assumptions except for the model for Mexico and Singapore. It failed the Independence of Residuals hypothesis test, however the Q-Q Plots said the opposite.

5.2 Overfitting

To check the overfitting for the models, we evaluated their training and test RMSE scores and concluded that none of the models were overfitted.

6. Challenges

- High variance and skewness in sales data.
- Shortcomings of Ad1 variable.
- Multiple time periods in datasets.

Appendix 1

US Mean Absolute Percentage Error (24.25%)

Time	Actual Unit Sold	Predicted Unit Sold
1	124.80	63.09

Singapore Mean Absolute Percentage Error (22.89%)

Time	Actual Unit Sold	Predicted Unit Sold
1	36.48	16.48

Spain Mean Absolute Percentage Error (26.52%)

Time	Actual Unit Sold	Predicted Unit Sold
1	15.63	22.33

Mexico Mean Absolute Percentage Error (21.43016%)

Time	Actual Unit Sold	Predicted Unit Sold
1	74.04	37.74

UK Mean Absolute Percentage Error (24.25%)

Time	Actual Unit Sold	Predicted Unit Sold
1	28.23	13.83

Overall MAPE is 23.29%

Appendix 2

1. Del Prete, M. and Samoggia, A. (2020) 'Chocolate consumption and purchasing behaviour review: Research issues and insights for future research', *Sustainability*, 12(14), p.5586. doi: 10.3390/su12145586.
2. Galanakis, C.M. ed. (2022) *Trends in Sustainable Chocolate Production*. Cham: Springer International Publishing. doi: https://doi.org/10.1007/978-3-030-90169-1.
3. KPMG (2014) *A taste of the future.* [pdf] Available at: https://assets.kpmg.com/content/dam/kpmg/pdf/2014/06/taste-of-the-future.pdf (Accessed: [date]).
4. Matejková, E. (2014) 'Analysis of consumer behavior at chocolate purchase', *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 62(6), pp.1511–1518. doi: 10.11118/actaun201462061511.
5. ReportLinker (2024) *Global Chocolate Market Report.* [online] Available at: https://www.reportlinker.com/clp/global/1878 (Accessed: [date]).
6. Sondhi, N. (2016) 'Segmenting and profiling the chocolate consumer: An emerging market perspective', *Journal of Food Products Marketing*, 22(3), pp.263–275. doi: 10.1080/10454446.2015.1048021.
7. Statista (2024) *Share of the global chocolate market by region.* [online] Available at: https://www-statista-com.manchester.idm.oclc.org/statistics/237744/share-of-the-global-chocolate-market-by-region (Accessed: 2024).