Price Optimization and Profitability Engine
Food Kiosk

Group 7 Authors:

- Yash Khare
- Manjusha Motamarry
- Margi Shah
- Mohan Bhosale
- Karthikeyan Sugavanan

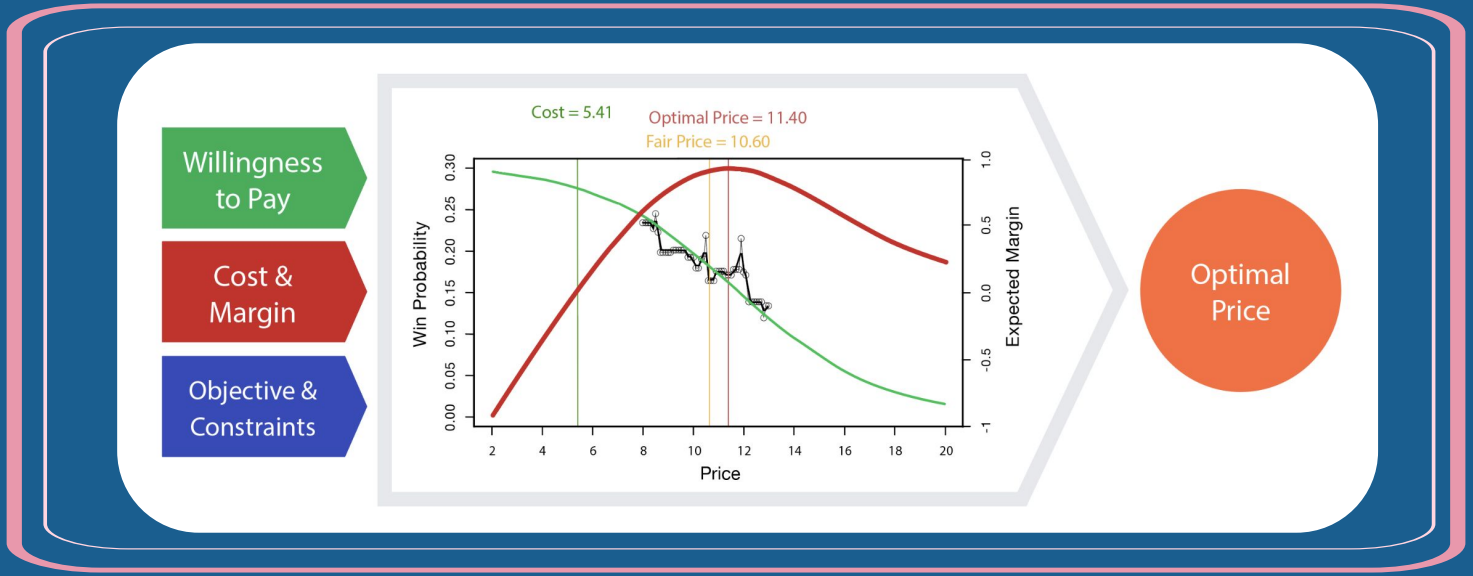
TABLE OF CONTENTS

UNDERSTANDING THE PROBLEM

FOOD KIOSK

This study quantifies the price elasticity of demand for food items at a Café within the Microsoft building in USA, aiming to determine optimal pricing adjustments in response to cost pressures while maintaining customer demand.

OBJECTIVES



- Understand Customer Response
- Determine Optimal Pricing
- Maximize Profit
- Leverage Sales Data for Pricing
- Diverse Data Utilization
- Focus on Price Elasticity

ELASTIC vs INELASTIC DEMAND

Price Elasticity: % change in quantity for 1 % increase in price

PED = ΔQ/ΔP

Elastic Demand

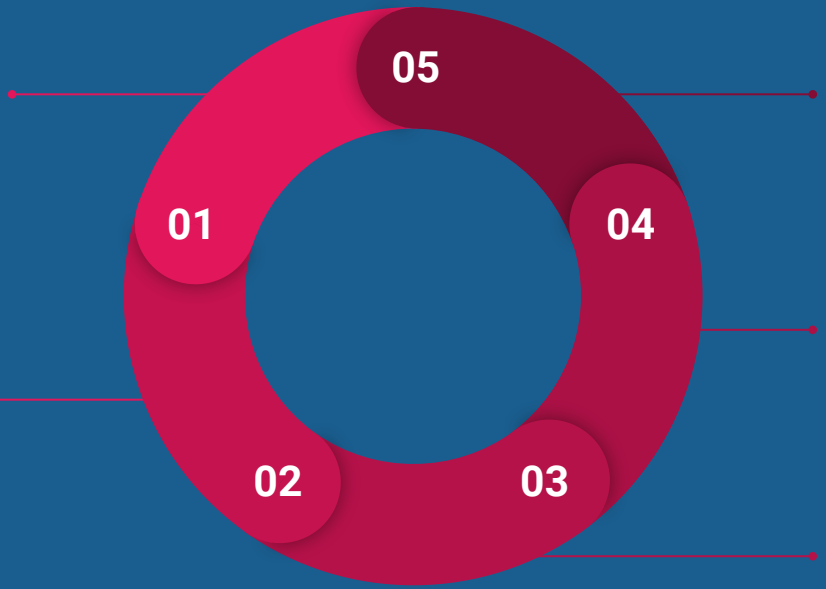
Inelastic Demand

Source:
https://demandeconomics.weebly.com/elastic-demand.html
https://demandeconomics.weebly.com/inelastic-demand.html

METHODOLOGY

Dataset Description: Data types, structure, and key columns

Initial EDA: Handling Null values, pivoting, and joining



Final Modeling: Multivariate demand predictions using R & Python

Feature Engineering and Feature Selection

Initial Modeling: Linear regression in r

FOOD KIOSK : SELL META-DATA

SELL_ID is the identifier of products that can be single items or a combo of items.

SELL_ID	SELL_CATEGORY	ITEM_ID	ITEM_NAME
1070	0	7821	BURGER
3055	0	3052	COFFEE
3067	0	5030	COKE
3028	0	6249	LEMONADE
2051	2	7821	BURGER
2051	2	5030	COKE
2052	2	7821	BURGER
2052	2	6249	LEMONADE
2053	2	7821	BURGER
2053	2	5030	COKE
2053	2	3052	COFFEE

FOOD KIOSK : TRANSACTION DATA

Each CALENDAR_DATE has four SELL_IDs with their PRICE and QUANTITY.

CALENDAR_DATE	PRICE	QUANTITY	SELL_ID	SELL_CATEGORY
2019-01-01	15.50	46	1070	0
2019-01-01	12.73	22	2051	2
2019-01-01	12.75	18	2052	2
2019-01-01	12.60	30	2053	2
2019-01-02	15.50	70	1070	0
2019-01-02	12.73	22	2051	2
2019-01-02	12.75	16	2052	2
2019-01-02	12.60	34	2053	2
2019-01-03	15.50	62	1070	0
2019-01-03	12.73	26	2051	2

FOOD KIOSK : DATE INFORMATION

CALENDAR_DATE	YEAR	HOLIDAY	IS_WEEKEND	IS_SCHOOLBREAK	AVERAGE_TEMPERATURE	IS_OUTDOOR
2019-02-02	2019	Luner New Year	0	0	32.0	1
2019-06-04	2019	Dragon Boat Festivel	0	0	78.8	1
2019-09-10	2019	Mid-Autumn Day	0	0	60.8	1
2019-10-01	2019	National Day	0	0	59.0	1
2019-04-30	2019	Labor Day	0	0	62.6	1
2019-04-03	2019	Qing Ming Festival	0	0	48.2	1
2022-09-03	2022	WWII Celebration	0	0	73.4	1
2019-01-01	2019	New Year	1	0	24.8	0

Each CALENDAR_DATE has external factors as columns

Replacing NULLs to get comprehensive data

CALENDAR_DATE <chr>	YEAR <dbl>	HOLIDAY <chr>
2019-01-01	2019	New Year
2019-01-02	2019	New Year
2019-01-03	2019	New Year
2019-01-04	2019	*NA*
2019-01-05	2019	*NA*
2019-01-06	2019	*NA*
2019-01-07	2019	*NA*
2019-01-08	2019	*NA*
2019-01-09	2019	*NA*
2019-01-10	2019	*NA*

CALENDAR_DATE <chr>	YEAR <dbl>	HOLIDAY <chr>
2019-01-01	2019	New Year
2019-01-02	2019	New Year
2019-01-03	2019	New Year
2019-01-04	2019	No Holiday
2019-01-05	2019	No Holiday
2019-01-06	2019	No Holiday
2019-01-07	2019	No Holiday
2019-01-08	2019	No Holiday
2019-01-09	2019	No Holiday
2019-01-10	2019	No Holiday

Pivot Wider

To have metadata information for a SELL_ID product in a single row.

SELL_ID <int>	SELL_CATEGORY <int>	ITEM_ID <int>	ITEM_NAME <chr>
1070	0	7821	BURGER
3055	0	3052	COFFEE
3067	0	5030	COKE
3028	0	6249	LEMONADE
2051	2	7821	BURGER
2051	2	5030	COKE
2052	2	7821	BURGER
2052	2	6249	LEMONADE
2053	2	7821	BURGER
2053	2	5030	COKE

SELL_ID <int>	BURGER <dbl>	COFFEE <dbl>	COKE <dbl>	LEMONADE <dbl>
1070	1	0	0	0
2051	1	0	1	0
2052	1	0	0	1
2053	1	1	1	0
3028	0	0	0	1
3055	0	1	0	0

Performing Joins on three tables

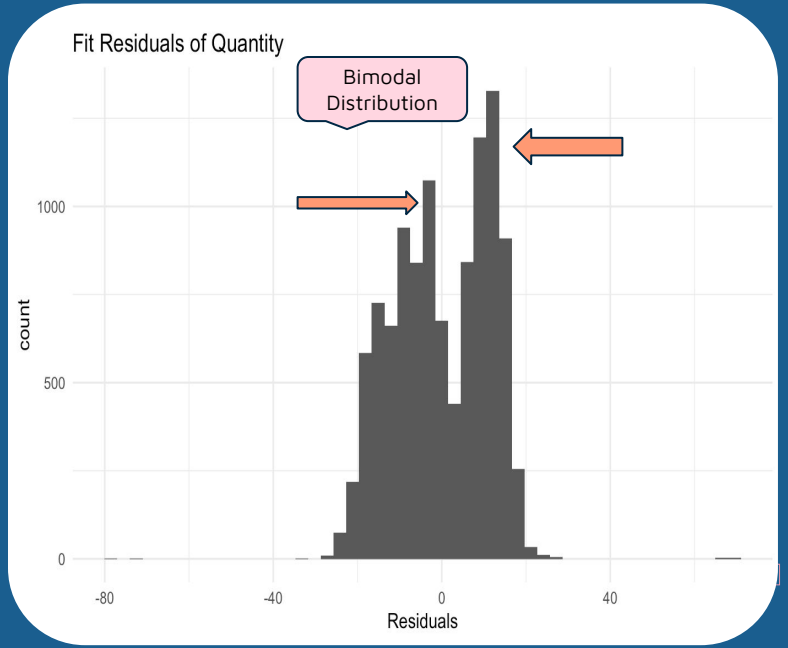
CALENDAR_DA... <chr>	PRICE <dbl>	QUANTITY <int>	SELL_ID <int>	SELL_CATEGORY <int>	BURGER <dbl>	COFFEE <dbl>	CO... <dbl>	LEMONA... <dbl>	YEAR <dbl>
2019-01-01	15.50	46	1070	0	1	0	0	0	2019
2019-01-01	12.73	22	2051	2	1	0	1	0	2019
2019-01-01	12.75	18	2052	2	1	0	0	1	2019
2019-01-01	12.60	30	2053	2	1	1	1	0	2019
2019-01-02	15.50	70	1070	0	1	0	0	0	2019
2019-01-02	12.73	22	2051	2	1	0	1	0	2019
2019-01-02	12.75	16	2052	2	1	0	0	1	2019
2019-01-02	12.60	34	2053	2	1	1	1	0	2019
2019-01-03	15.50	62	1070	0	1	0	0	0	2019
2019-01-03	12.73	26	2051	2	1	0	1	0	2019

Inner join on **sold** and **transaction** on **sell_id**.
And **Inner Join** on above **merged** data with **dates** on **calendar_date .**

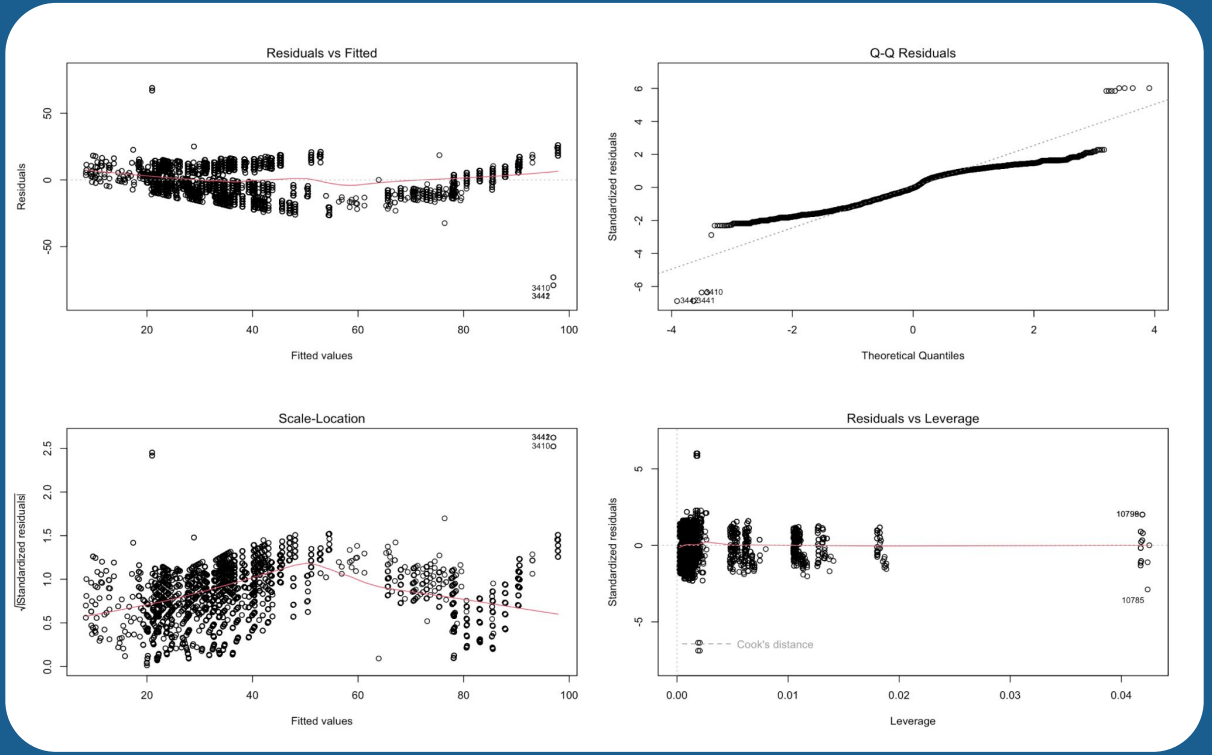
For a comprehensive analysis where we might need to study transaction details alongside the sales data.

Fitting Initial Linear Regression Model gives Bimodal Distribution of Residuals

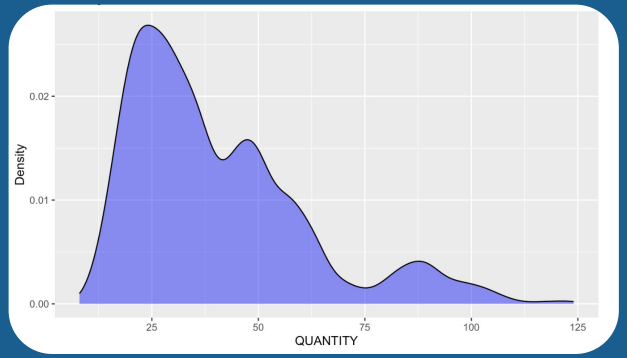
```
Call:
lm(formula = QUANTITY ~ PRICE + SELL_ID + CALENDAR_DATE + HOLIDAY +
    IS_WEEKEND + IS_SCHOOLBREAK + IS_OUTDOOR + AVERAGE_TEMPERATURE,
    data = merged_data)

Residuals:
    Min      1Q  Median      3Q     Max
-79.008  -9.117  -0.426  10.213  69.025

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)               2.218e+02  6.563e+00   33.800   <2e-16 ***
PRICE                    -4.858e+00  1.311e-01  -37.056   <2e-16 ***
SELL_ID                  -6.333e-02  5.329e-04 -118.832   <2e-16 ***
CALENDAR_DATE             7.174e-06  3.033e-04    0.024    0.981
HOLIDAYLabor Day          5.135e-01  1.660e+00    0.309    0.757
HOLIDAYLuner New Year    -4.494e-01  1.460e+00   -0.308    0.758
HOLIDAYMid-Autumn Day     7.899e-01  1.932e+00    0.409    0.683
HOLIDAYNational Day      -4.011e-02  1.472e+00   -0.027    0.978
HOLIDAYNew Year          -1.048e+00  1.767e+00   -0.593    0.553
HOLIDAYNo Holiday         1.174e+01  1.187e+00    9.890   <2e-16 ***
HOLIDAYQing Ming Festival 9.689e-01  1.666e+00    0.582    0.561
HOLIDAYWWII Celebration   2.408e-01  2.628e+00    0.092    0.927
IS_WEEKEND               -1.250e+01  2.449e-01  -51.048   <2e-16 ***
IS_SCHOOLBREAK           -7.142e-02  3.181e-01   -0.224    0.822
IS_OUTDOOR               -7.486e+00  4.005e-01  -18.691   <2e-16 ***
AVERAGE_TEMPERATURE      -1.630e-03  7.178e-03   -0.227    0.820
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.48 on 10824 degrees of freedom
Multiple R-squared:  0.6995,    Adjusted R-squared:  0.6991
F-statistic:  1680 on 15 and 10824 DF,  p-value: < 2.2e-16
```

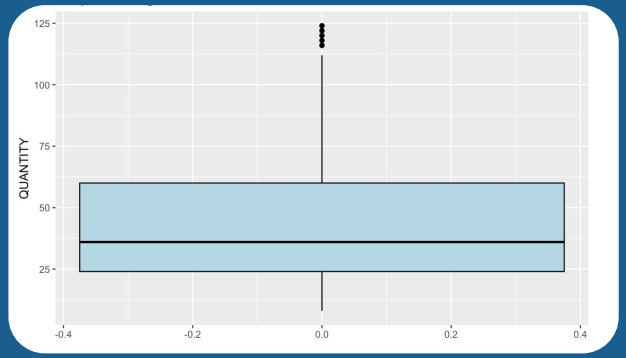


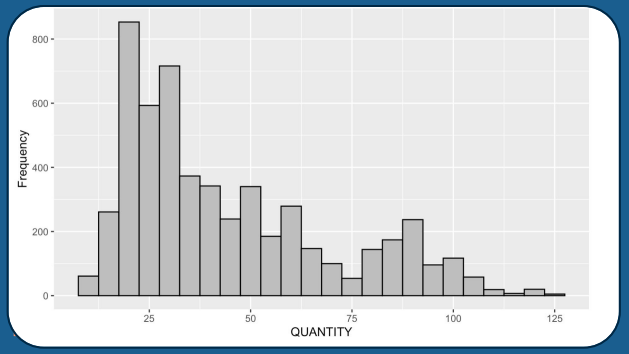
Residual Analysis suggests Heteroscedasticity

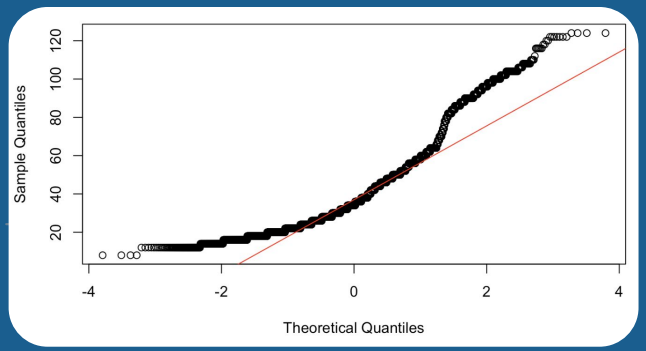


Quantity is Right Skewed. Most of the items quantity is between 25-50 units



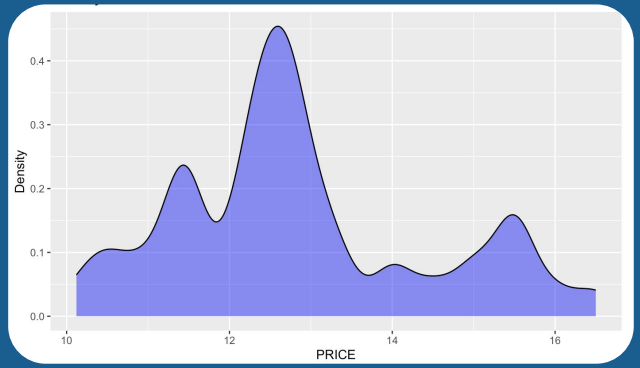


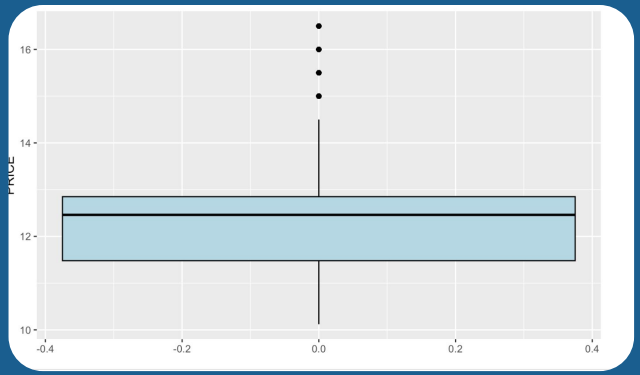


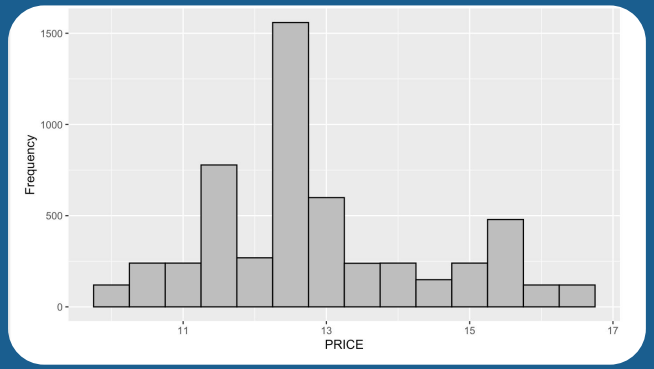


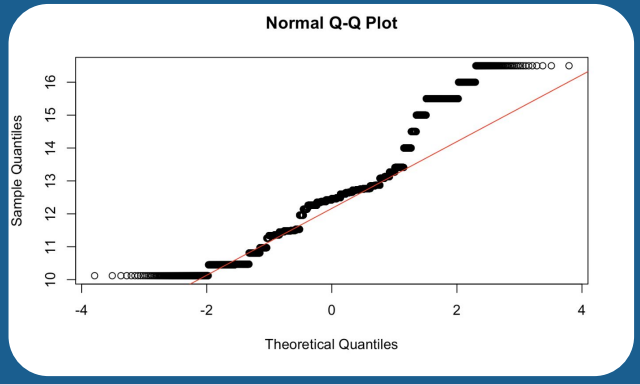
Exploring Numerical features : Quantity

Price has a Multimodal Distribution. Most of the item's price lies between 12-13$



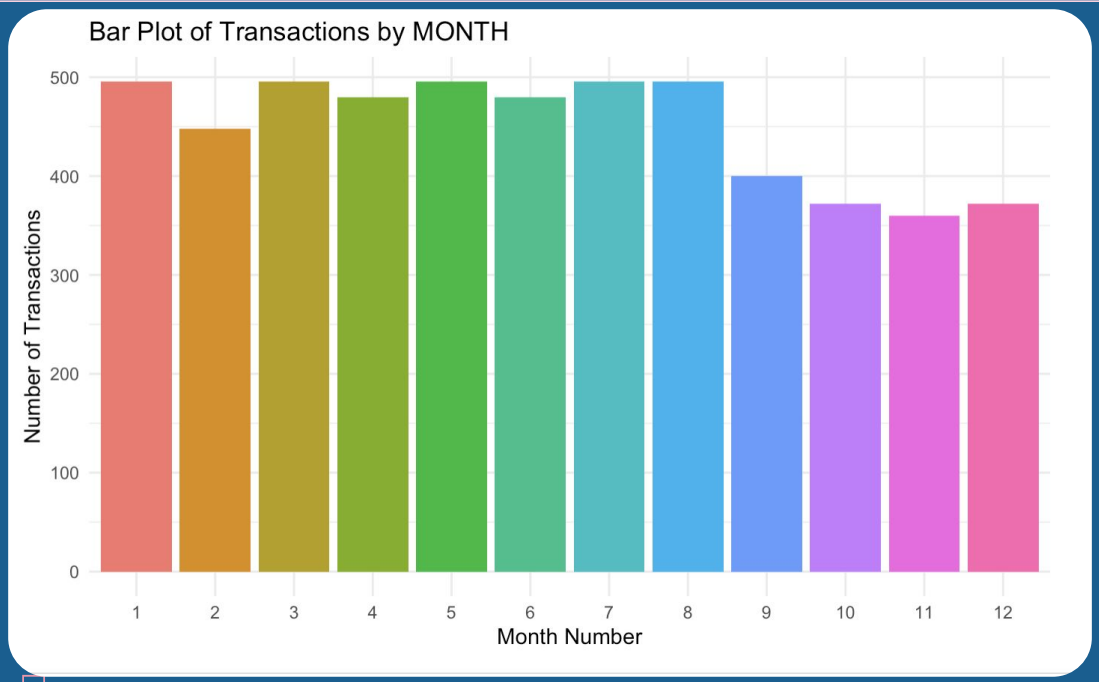






Exploring Numerical Features : Price

Number of Transactions are comparatively less from September to December

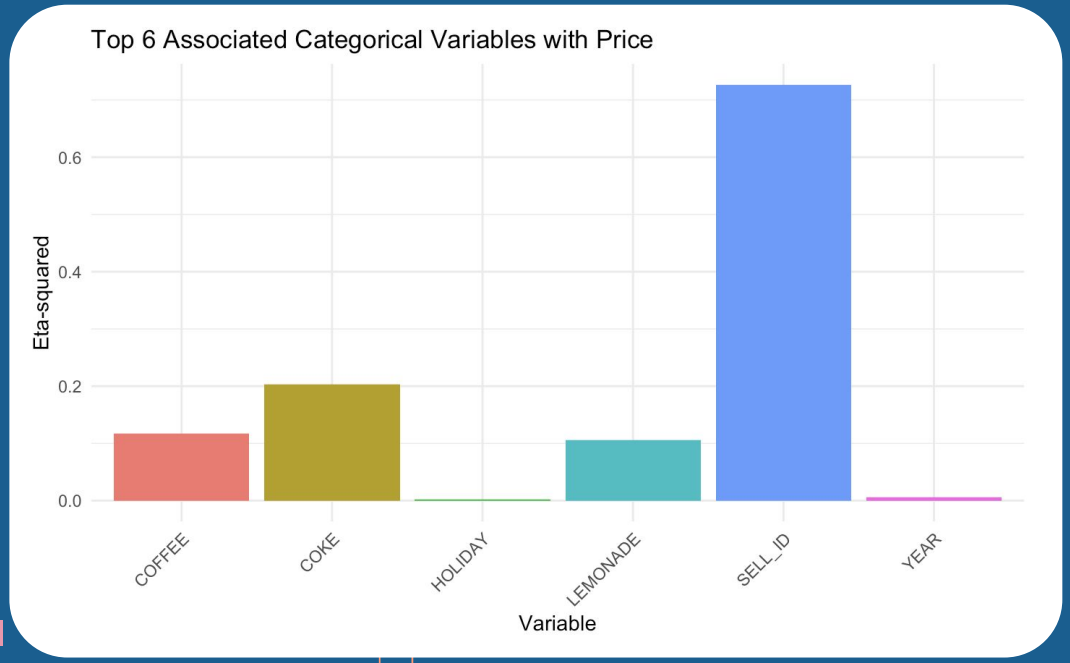


Exploring Categorical Variable:
Date

Sell_id is related to Price the most. Holiday is least related to price

Eta-squared values:

SELL_ID	COFFEE	COKE	LEMONADE	YEAR	HOLIDAY	IS_WEEKEND	IS_OUTDOOR
7.269516e-01	1.169177e-01	2.029792e-01	1.057703e-01	6.245940e-03	1.671734e-03	7.480624e-07	1.340378e-03

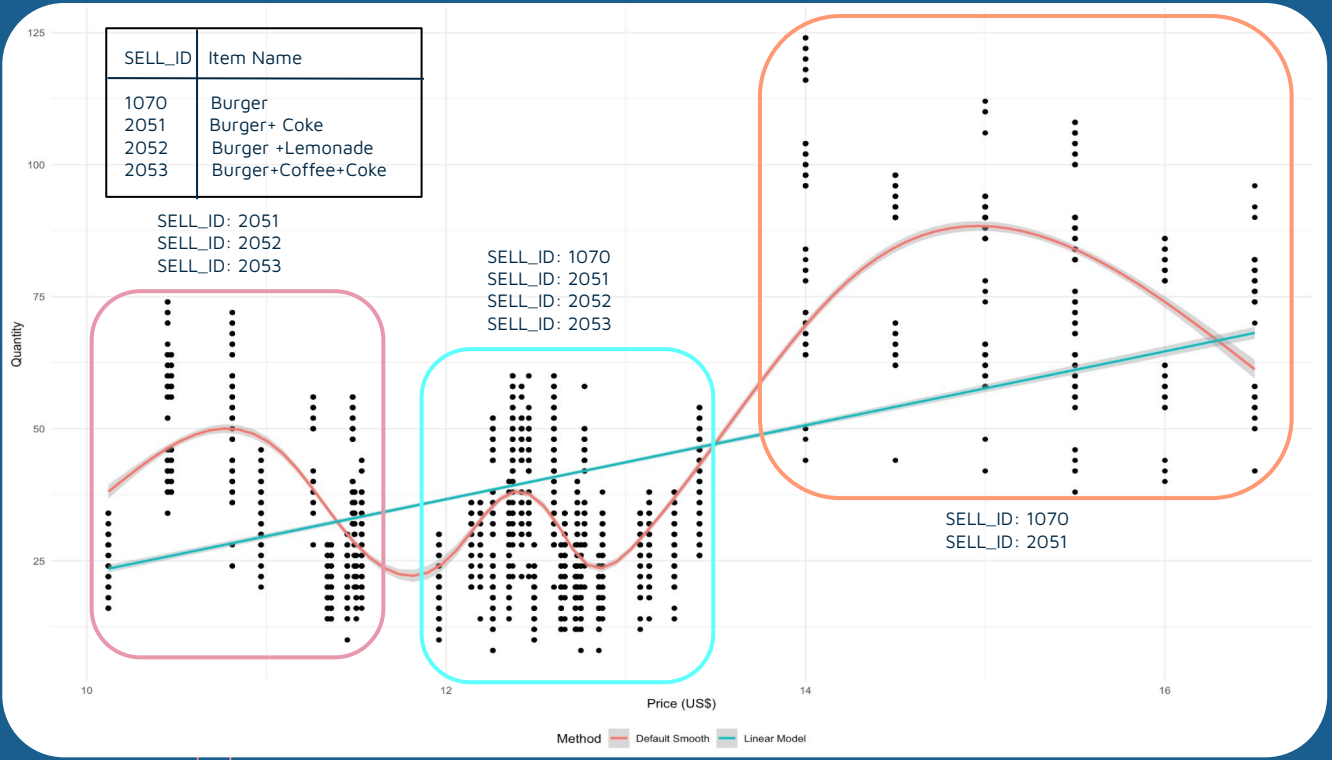


Eta-squared values tells how much of the variation in price can be explained by each of the categorical variables.

Multivariate Analysis

Different correlations for different groups of sell_id

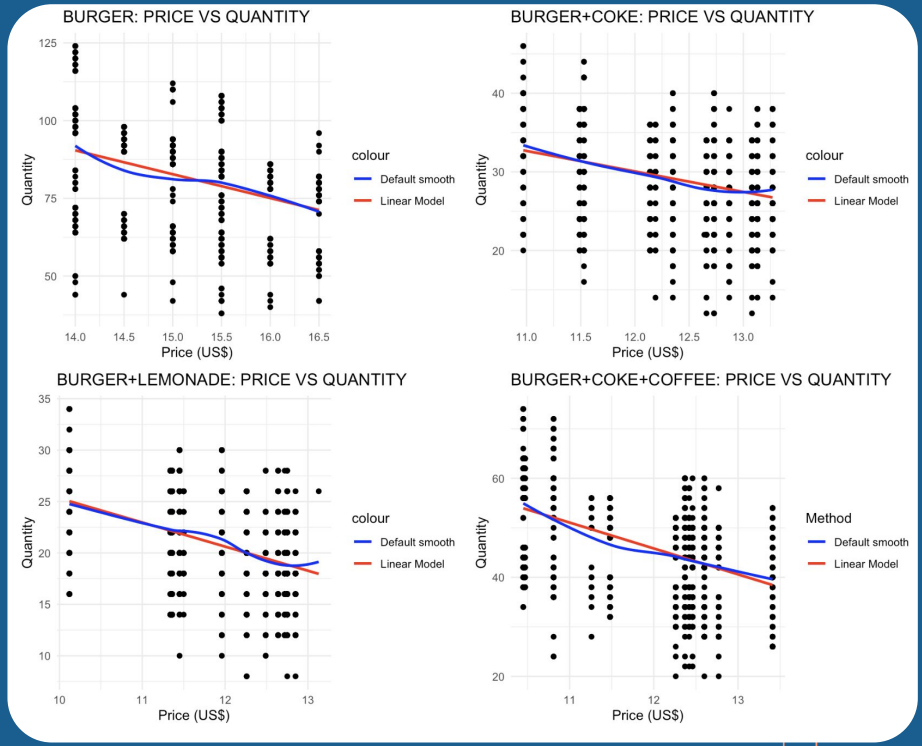
SELL_ID	Item Name
1070	Burger
2051	Burger+ Coke
2052	Burger +Lemonade
2053	Burger+Coffee+Coke



Price Vs QUANTITY

Negative correlation across all individual Items

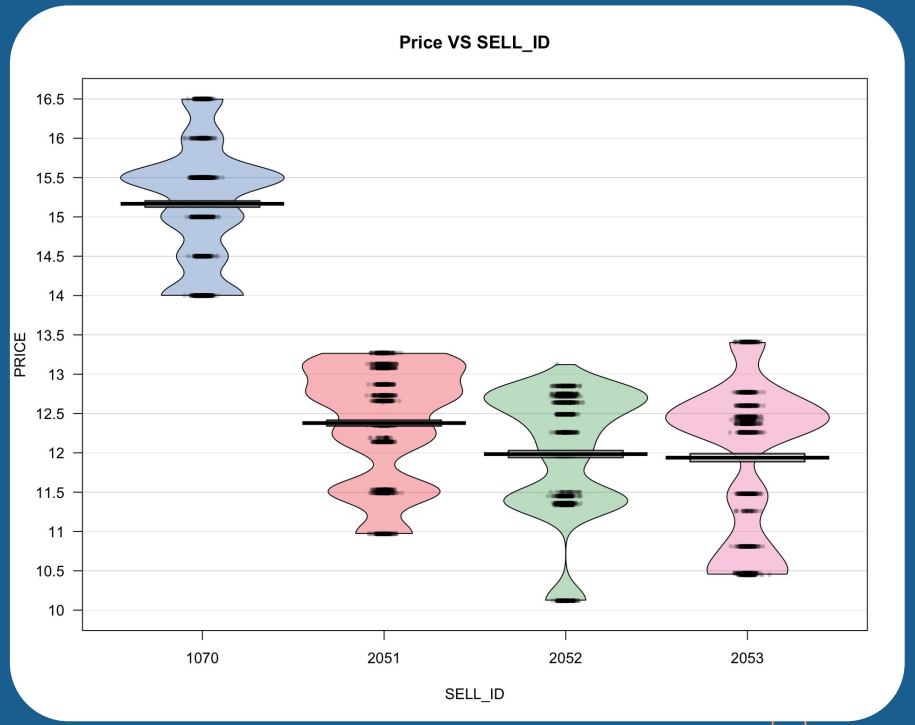
The negative relationship across all plots is consistent with basic economic principle: as **prices increase, demand typically decreases**.



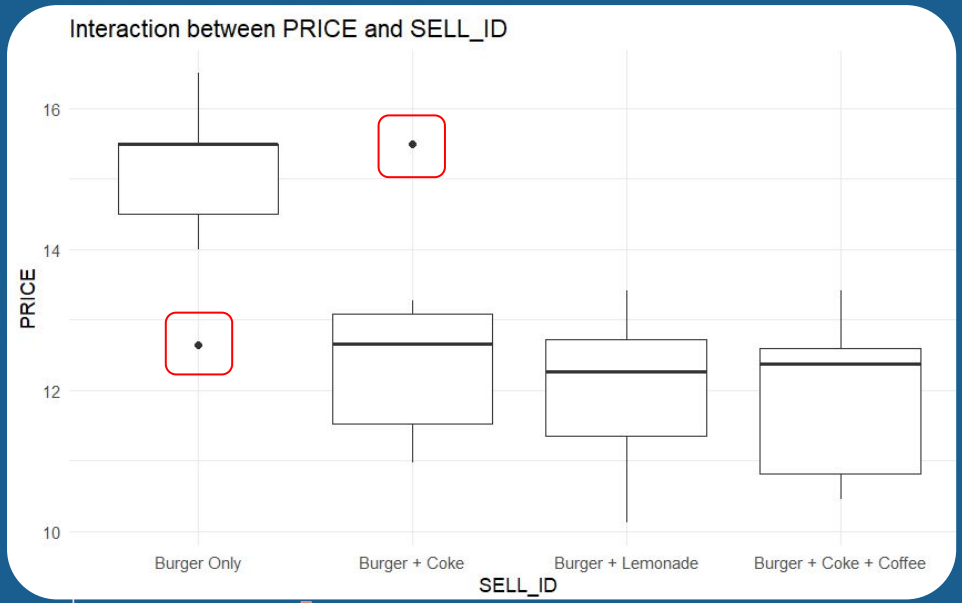
Burger(SELL_ID=1070) has the highest median than others

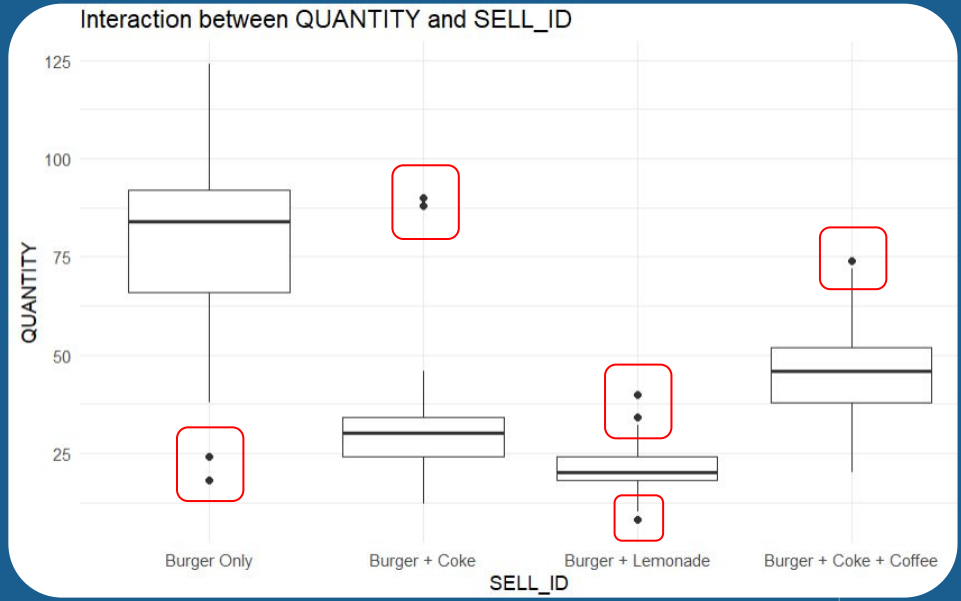
Price Vs SELL_ID

SELL_ID	Item Name
1070	Burger
2051	Burger+ Coke
2052	Burger + Lemonade
2053	Burger+ Coffee+ Coke



Understanding Outliers Between Price and Quantity among different SELL_ID





Finding Outliers

CALENDAR_DATE <chr>	PRICE <dbl>	QUANTITY <int>	SELL_ID <fctr>
2020-03-01	12.64	24	1070
2020-03-01	12.64	24	1070
2020-03-01	12.64	18	1070
2020-03-01	12.64	18	1070

Table 1

CALENDAR_DATE <chr>	PRICE <dbl>	QUANTITY <int>	SELL_ID <fctr>
2020-02-27	15.50	90	1070
2020-02-28	15.50	84	1070
2020-03-01	15.50	90	1070
2020-03-01	15.50	90	1070
2020-03-01	12.64	24	1070
2020-03-01	12.64	24	1070
2020-03-01	15.50	90	1070
2020-03-01	15.50	90	1070
2020-03-01	12.64	18	1070
2020-03-01	12.64	18	1070

Table 2

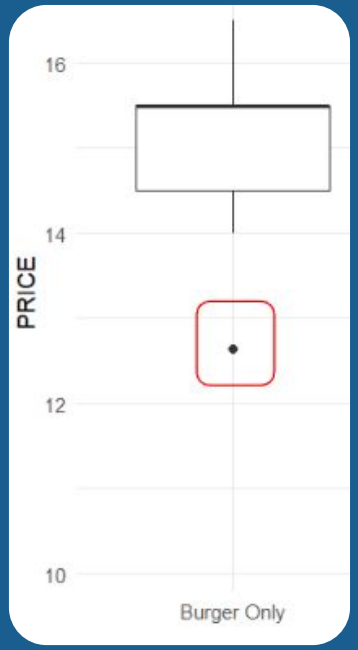


Fig 1

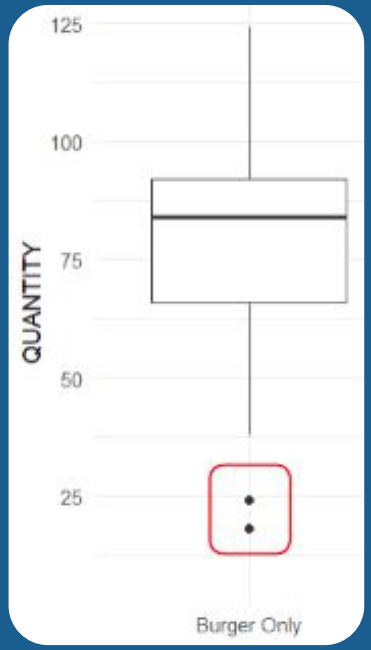


Fig 2

Filtering data points with Price less than 14 reveals multiple rows with the same date, potentially explaining outliers in the Quantity vs Sell_ID

Replacing Duplicates

CALENDAR_DATE <chr>	PRICE <dbl>	QUANTITY <int>	SELL_ID <int>	SELL_CATEGORY <int>	BURGER <dbl>	COFFEE <chr>	CO... <chr>	LEMONA... <chr>	YEAR <dbl>
2020-03-01	15.5	90	1070	0	1	0	0	0	2020

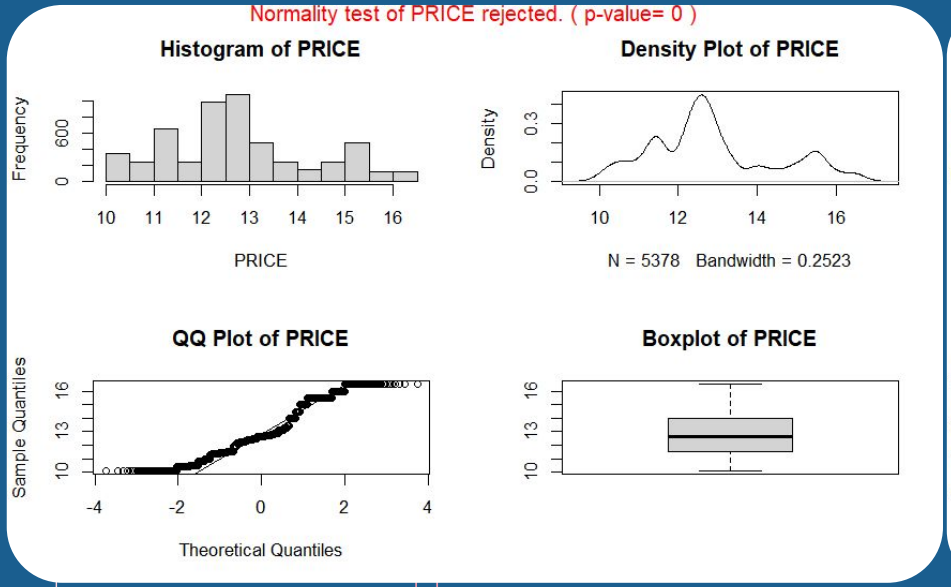
CALENDAR_DATE <chr>	SELL_ID <int>	n <int>
2020-03-01	2051	8
2020-03-01	2052	8
2020-03-01	2053	8

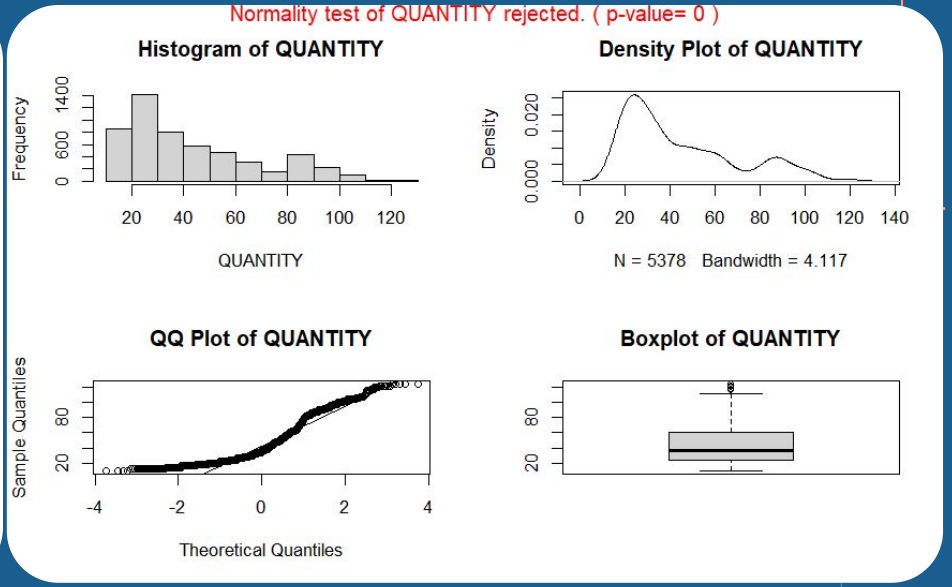
CALENDAR_DATE <chr>	PRICE <dbl>	QUANTITY <int>	SELL_ID <fctr>	SELL_CATEGORY <int>	BURGER <fctr>	COFFEE <fctr>	CO... <fctr>	LEMONA... <fctr>	YEAR <fctr>
2020-03-01	15.50	90	1070	0	1	0	0	0	2020
2020-03-01	12.64	22	2051	2	1	0	1	0	2020
2020-03-01	13.13	26	2052	2	1	0	0	1	2020
2020-03-01	13.41	40	2053	2	1	1	1	0	2020

Replacing duplicate rows with cleaned_data row and checking for other similar outliers in other Sell_ids'.

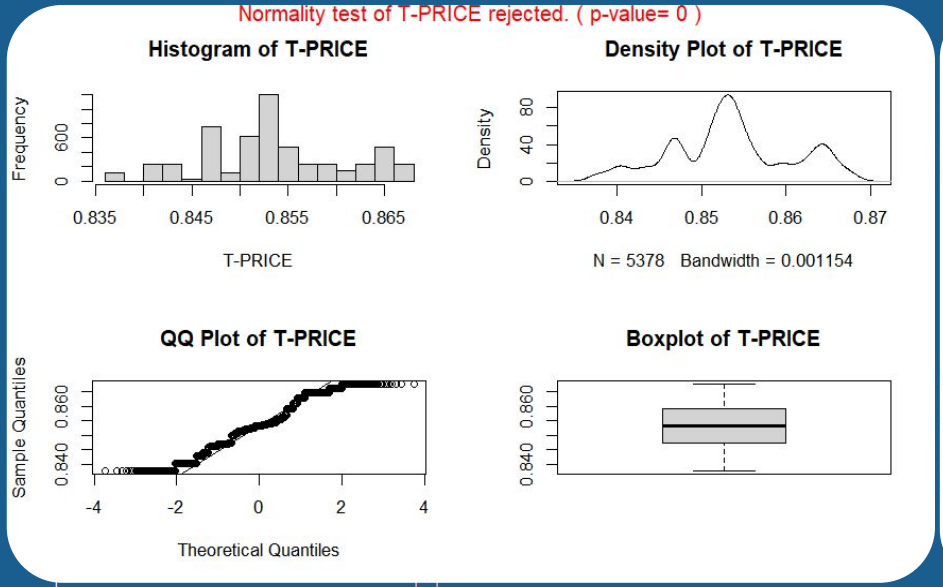
Rejecting Null Hypothesis that distribution is normal

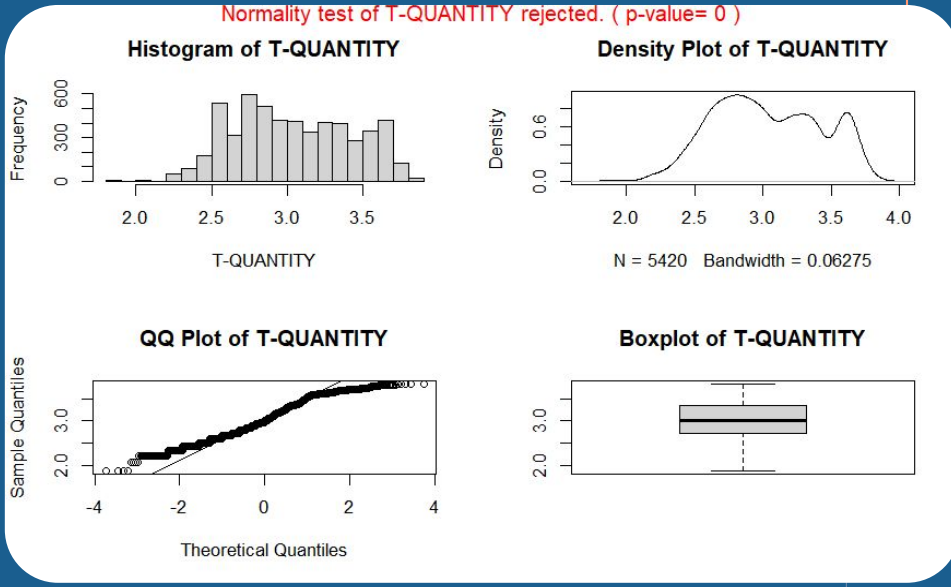
- Shapiro-Wilk normality test



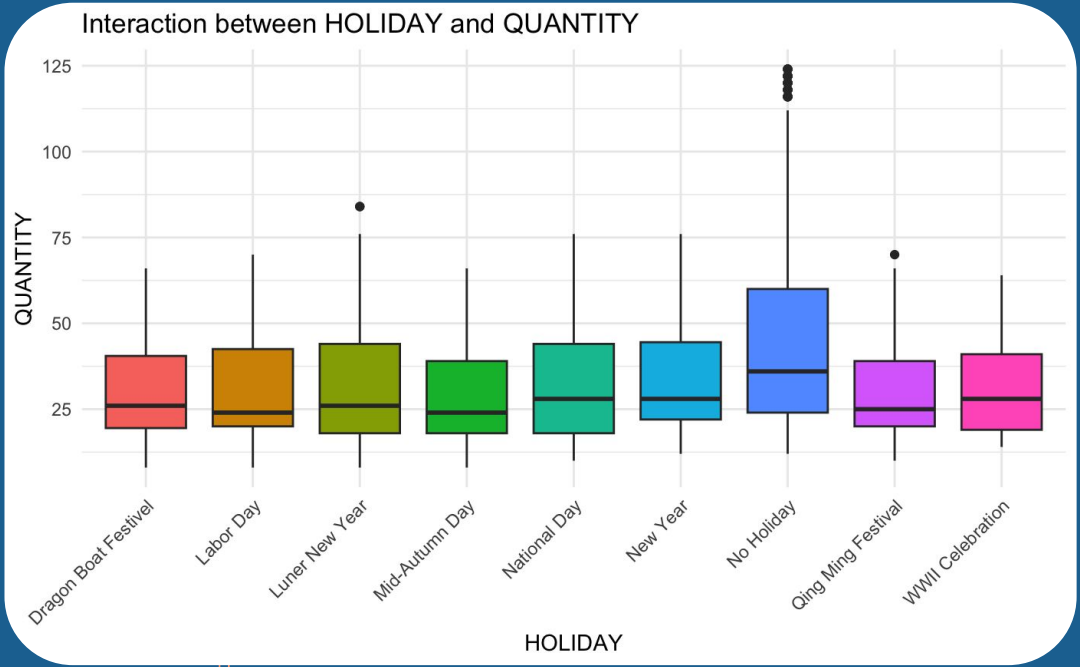


Box-Cox Transformation reduces skewness





Integrating Holiday Indicator Variable

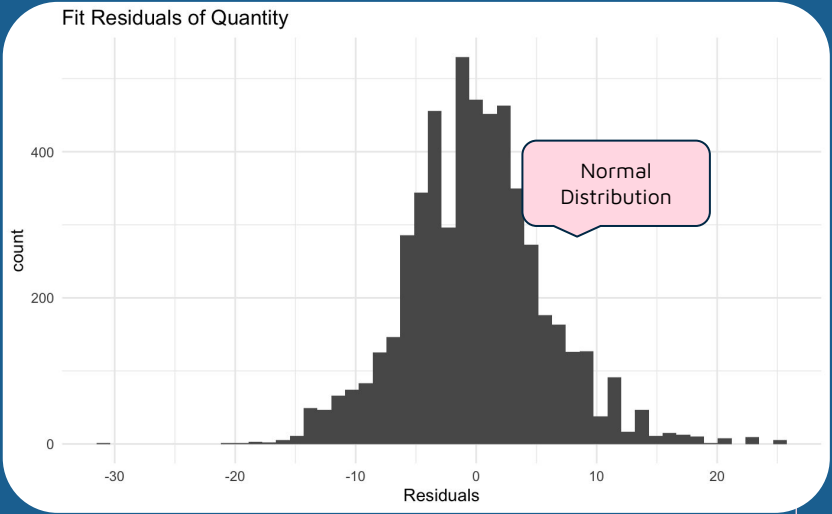


CALENDAR_DATE	HOLIDAY	IS_HOLIDAY
2019-01-01	New Year	1
2019-02-02	Luner New Year	1
2019-04-03	Qing Ming Festival	1
2019-04-30	Labor Day	1
2019-06-04	Dragon Boat Festivel	1
2019-09-10	Mid-Autumn Day	1
2019-10-01	National Day	1
2022-09-03	WWII Celebration	1

Creating new variable IS_HOLIDAY , for ML to find patterns

Feature Engineering results in Normal distribution of residuals

```
Call:
lm(formula = QUANTITY ~ PRICE + SELL_ID + HOLIDAY + IS_WEEKEND +
    IS_OUTDOOR, data = kiosk_data)

Residuals:
     Min       1Q   Median       3Q      Max
-31.4025  -3.6286  -0.0382   3.2027  24.7459

Coefficients:
                              Estimate Std. Error  t value Pr(>|t|)
(Intercept)                1.511e+02  1.754e+00   86.144   <2e-16 ***
PRICE                     -4.634e+00  9.884e-02  -46.883   <2e-16 ***
SELL_ID2051               -6.533e+01  3.558e-01 -183.654   <2e-16 ***
SELL_ID2052               -7.557e+01  3.868e-01 -195.399   <2e-16 ***
SELL_ID2053               -5.029e+01  3.904e-01 -128.822   <2e-16 ***
HOLIDAYLabor Day           4.228e-01  1.192e+00    0.355    0.723
HOLIDAYLuner New Year     -9.821e-01  1.016e+00   -0.967    0.334
HOLIDAYMid-Autumn Day      6.436e-01  1.389e+00    0.463    0.643
HOLIDAYNational Day       -1.492e-01  1.057e+00   -0.141    0.888
HOLIDAYNew Year           -1.776e+00  1.262e+00   -1.407    0.159
HOLIDAYNo Holiday          1.303e+01  8.476e-01   15.378   <2e-16 ***
HOLIDAYQing Ming Festival  9.101e-01  1.192e+00    0.763    0.445
HOLIDAYWWII Celebration   -3.907e-04  1.886e+00    0.000    1.000
IS_WEEKEND1               -1.402e+01  1.763e-01  -79.544   <2e-16 ***
IS_OUTDOOR1               -8.500e+00  2.376e-01  -35.776   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.84 on 5377 degrees of freedom
Multiple R-squared:  0.9477,    Adjusted R-squared:  0.9475
F-statistic:  6956 on 14 and 5377 DF,  p-value: < 2.2e-16
```



Normal distribution of residual, states that it’s more likely that the model correctly captures the linear relationship without missing any nonlinear patterns

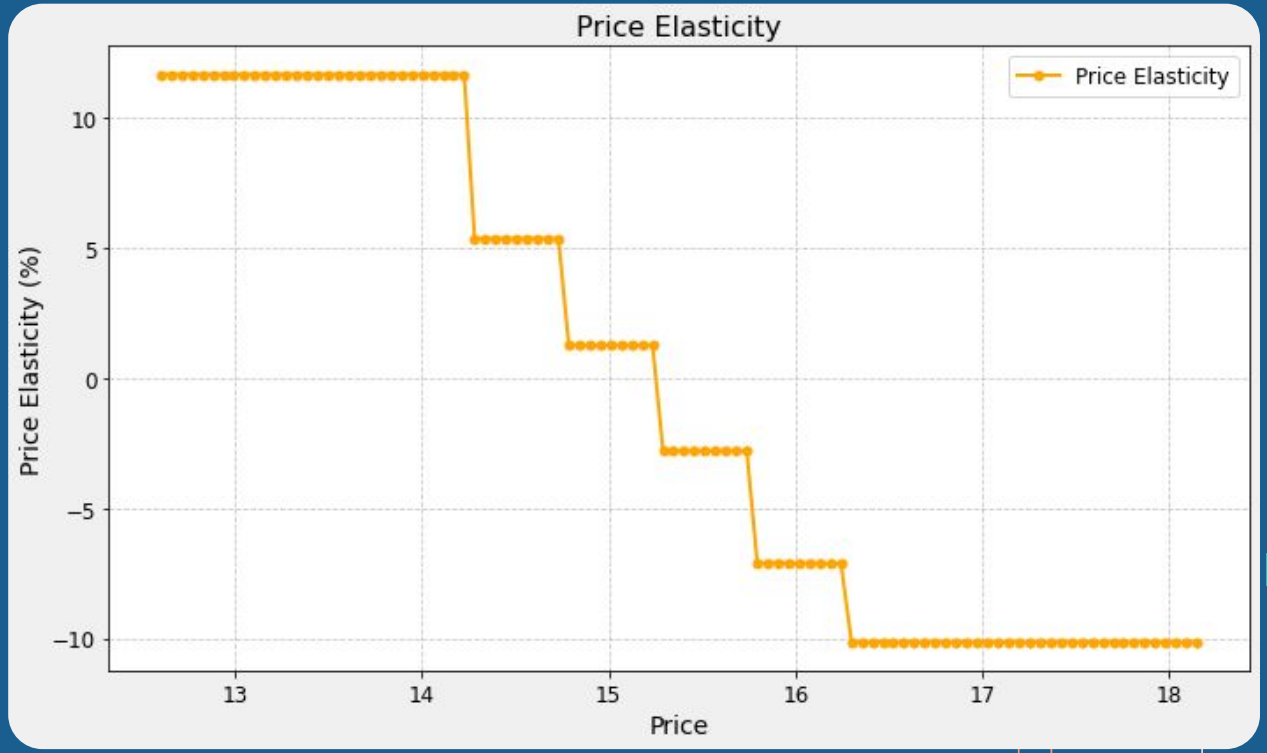
Choosing an Algorithm: GB Gives the best test results

	Train R2 Scores	Train RMSE Scores	Test R2 Scores	Test RMSE Scores
AdaBoost	0.954115	5.505732	0.950299	5.488622
GradientBoosting	0.987480	2.875976	0.986042	2.908713
XGBoost	0.995321	1.758148	0.983585	3.154342
LightGBM	0.990218	2.542050	0.985694	2.944727
CatBoost	0.991068	2.429089	0.985873	2.926177

A Gradient Boosting algorithm is more suitable for our data as it performs very good on the test dataset

Price Elasticity: Change in price significantly impact the demand

plot suggests different price thresholds at which the quantity demanded responds differently to price changes. This type of analysis is crucial for determining optimal pricing strategies.



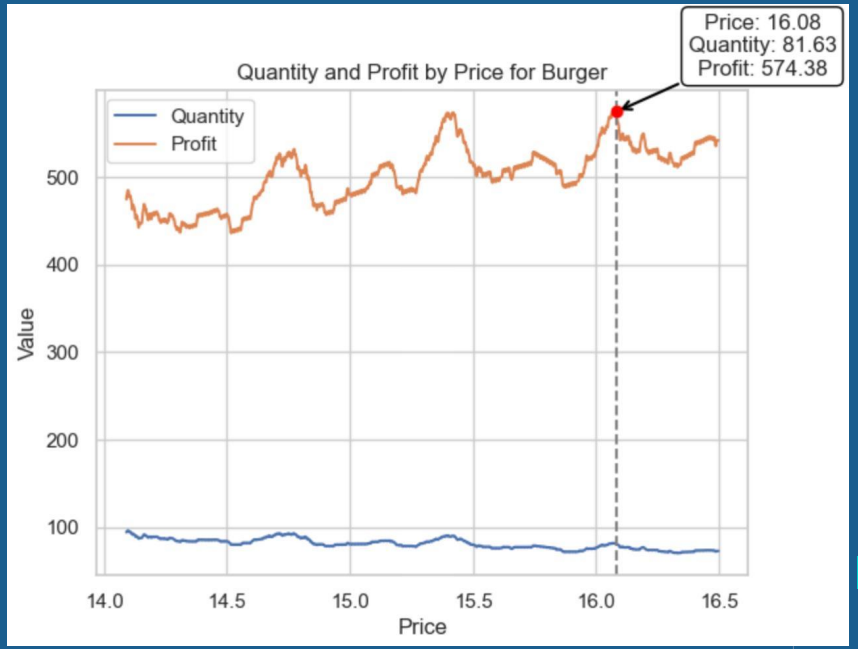
Predicting ‘QUANTITY’ keeping other variables constant

PRICE	QUANTITY	IS_WEEKEND	IS_OUTDOOR	IS_SCHOOLBREAK	IS_HOLIDAY
14.000000	0	1	0	0	1
14.001855	0	0	0	0	1
14.003709	0	0	1	0	1
14.005564	0	0	1	0	0
14.007418	0	0	0	0	0
14.009273	0	0	0	0	0
14.011128	0	1	0	0	0
14.012982	0	1	0	0	0
14.014837	0	0	0	0	0
14.016691	0	0	0	0	0
14.018546	0	0	0	0	0
14.020401	0	0	0	0	0
14.022255	0	0	0	0	0
14.024110	0	1	0	0	0
14.025964	0	1	1	0	0
14.027819	0	0	0	0	0
14.029674	0	0	0	0	0
14.031528	0	0	0	0	0
14.033383	0	0	0	0	0
14.035237	0	0	0	0	0

PRICE	QUANTITY	IS_WEEKEND	IS_OUTDOOR	IS_SCHOOLBREAK	IS_HOLIDAY
14.000000	55.549296	1	0	0	1
14.001855	81.974496	0	0	0	1
14.003709	68.039698	0	1	0	1
14.005564	100.248964	0	1	0	0
14.007418	120.470757	0	0	0	0
14.009273	119.318793	0	0	0	0
14.011128	80.931683	1	0	0	0
14.012982	80.595575	1	0	0	0
14.014837	120.687895	0	0	0	0
14.016691	120.687895	0	0	0	0
14.018546	120.937744	0	0	0	0
14.020401	120.829389	0	0	0	0
14.022255	119.992311	0	0	0	0
14.024110	81.334463	1	0	0	0
14.025964	68.433908	1	1	0	0
14.027819	120.701193	0	0	0	0
14.029674	120.701193	0	0	0	0
14.031528	121.261369	0	0	0	0
14.033383	120.715227	0	0	0	0
14.035237	120.250289	0	0	0	0

Predicting QUANTITY values keeping other variables constant as per the elasticity definition

16.08 is the Optimum Price to gain the maximum profit for the Burger.



Getting an OPTIMAL Price for different items which would result in the overall profit maximization.

Model Comparison: Separate modeling yield better results

Programming Language	Model Name	# Predictors	Items	RMSE	R-Squared
R	Linear Model	8	Burger + Coke + Coffee + Lemonade	11.48	69.95%
R	Linear Model with feature engineering	5	Burger + Coke + Coffee + Lemonade	5.84	94.77%
Python	Gradient Boosting	7	Burger	3.26	96.03%
Python	Gradient Boosting	7	Coke	3.11	71.89%
Python	Gradient Boosting	7	Coffee	2.92	91.24%
Python	Gradient Boosting	7	Lemonade	2.75	62.64%

Extending Price Optimization to any Business Project

Our model's versatility allows seamless integration into diverse industries. It will help to gain insights into customer behavior and market dynamics.

Empower business to make informed "Pricing Decisions".

THANK YOU