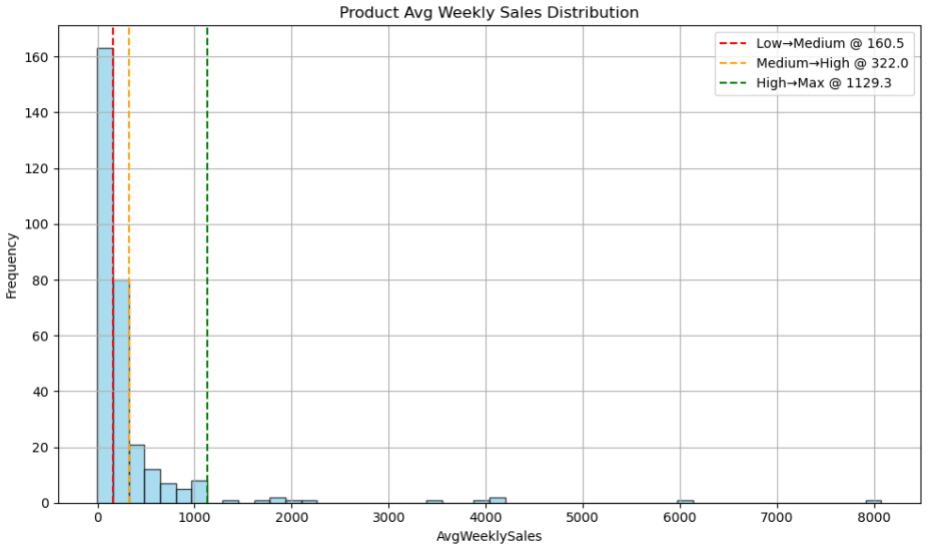
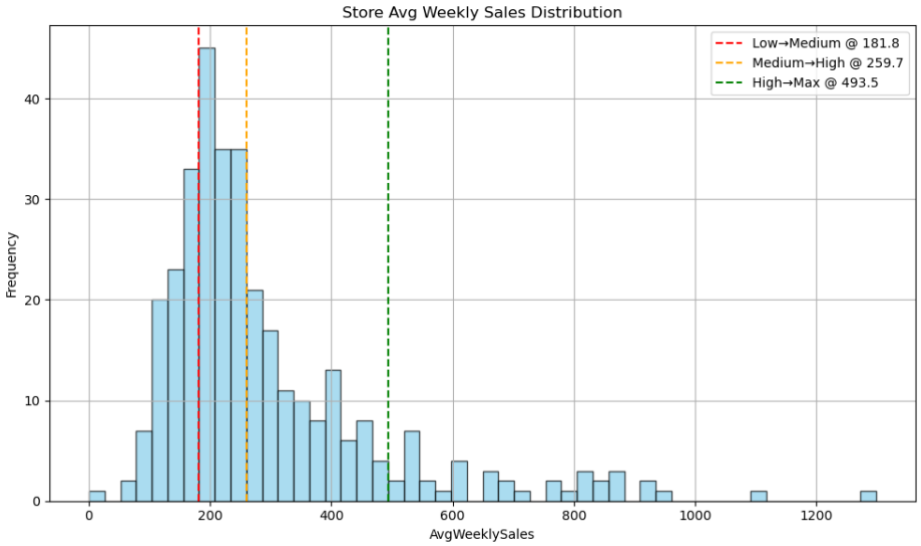
A.

a. What are your criteria for separating Fast, Medium, and Slow items? Why?

I used the cumulative distribution function (CDF) of average weekly sales per product during non-promotion periods. The gradient of the CDF was calculated and used to identify cutoff points:

- Fast items are those in the tail of the CDF where the gradient flattens, indicating high weekly sales.
- Medium items are in the transition region, where the gradient decreases after peaking.
- Slow items are in the region with the steepest gradient, corresponding to low weekly sales.





This gradient-based method captures natural clusters in the distribution without relying on arbitrary thresholds. It helps reflect the actual sales dynamics and consumer demand levels.

b. What are your criteria for separating Fast, Medium, and Slow stores? Why?

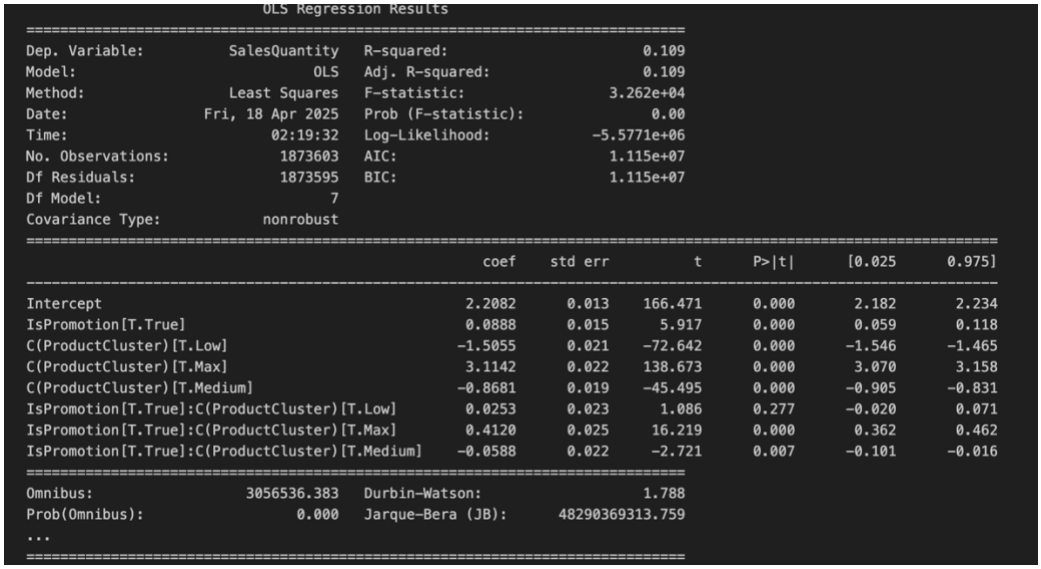
The same gradient method was used for stores. The average weekly sales for each store are calculated, and the distribution of these values was analyzed.

- **Fast stores** are in the flatter, high-value tail of the CDF.
- **Medium stores** lie in the mid-gradient zone.
- **Slow stores** are clustered in the high-gradient, low-sales region.

Using a consistent, data-driven segmentation for both items and stores improves interpretability and avoids biases. This also helps analyze how promotion responses vary by store performance levels.

c. Which items experienced the biggest sale increase during promotions?

Based on the regression results (model_both), Fast items sold the highest uplift during promotions. This is reflected by the significant and positive interaction coefficient for Fast items (**+0.4216**, $p < 0.001$). This indicates that promotions had a strong amplifying effect on already popular items, suggesting that the promotion reinforced existing customer interest.



OLS Regression Results

Dep. Variable:	SalesQuantity	R-squared:	0.109
Model:	OLS	Adj. R-squared:	0.109
Method:	Least Squares	F-statistic:	3.262e+04
Date:	Fri, 18 Apr 2025	Prob (F-statistic):	0.00
Time:	02:19:32	Log-Likelihood:	-5.5771e+06
No. Observations:	1873603	AIC:	1.115e+07
Df Residuals:	1873595	BIC:	1.115e+07
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.2082	0.013	166.471	0.000	2.182	2.234
IsPromotion[T.True]	0.0888	0.015	5.917	0.000	0.059	0.118
C(ProductCluster)[T.Low]	-1.5055	0.021	-72.642	0.000	-1.546	-1.465
C(ProductCluster)[T.Max]	3.1142	0.022	138.673	0.000	3.070	3.158
C(ProductCluster)[T.Medium]	-0.8681	0.019	-45.495	0.000	-0.905	-0.831
IsPromotion[T.True]:C(ProductCluster)[T.Low]	0.0253	0.023	1.086	0.277	-0.020	0.071
IsPromotion[T.True]:C(ProductCluster)[T.Max]	0.4120	0.025	16.219	0.000	0.362	0.462
IsPromotion[T.True]:C(ProductCluster)[T.Medium]	-0.0588	0.022	-2.721	0.007	-0.101	-0.016

Omnibus:	3056536.383	Durbin-Watson:	1.788
Prob(Omnibus):	0.000	Jarque-Bera (JB):	48290369313.759

...

d. Are there stores that have higher promotion reaction?

Yes. The regression coefficient for promotion interaction with Fast stores is **+0.6870**, with a very high t-statistic (33.139) and $p < 0.001$. In contrast, for Slow stores, the coefficient is negative **(-0.0549**, $p = 0.040$**).** This shows that **Fast stores** respond much more positively to promotions than Slow stores.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          SalesQuantity   R-squared:                       0.008
Model:                            OLS   Adj. R-squared:                  0.008
Method:                 Least Squares   F-statistic:                     2082.
Date:                Fri, 18 Apr 2025   Prob (F-statistic):               0.00
Time:                        02:19:37   Log-Likelihood:            -5.6775e+06
No. Observations:             1873572   AIC:                         1.135e+07
Df Residuals:                 1873564   BIC:                         1.136e+07
Df Model:                           7
Covariance Type:            nonrobust
================================================================================================================
                                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
Intercept                                       2.1926      0.014    161.906      0.000       2.166       2.219
IsPromotion[T.True]                             0.1462      0.015      9.583      0.000       0.116       0.176
C(StoreCluster)[T.Low]                         -0.5722      0.024    -24.008      0.000      -0.619      -0.526
C(StoreCluster)[T.Max]                          0.6646      0.022     30.260      0.000       0.622       0.708
C(StoreCluster)[T.Medium]                      -0.3027      0.020    -15.264      0.000      -0.342      -0.264
IsPromotion[T.True]:C(StoreCluster)[T.Low]     -0.0549      0.027     -2.051      0.040      -0.107      -0.002
IsPromotion[T.True]:C(StoreCluster)[T.Max]      0.0760      0.025      3.046      0.002       0.027       0.125
IsPromotion[T.True]:C(StoreCluster)[T.Medium]  -0.0291      0.022     -1.302      0.193      -0.073       0.015
==============================================================================
Omnibus:                  2987816.466   Durbin-Watson:                   1.597
Prob(Omnibus):                  0.000   Jarque-Bera (JB):      33118186080.410
...
```

e. What is the biggest effect explaining sales change during promotions?

The **product cluster** has the strongest effect:

- Regression R^2 for product model: **0.109**
- Regression R^2 for store model: **0.008**
- Combined model R^2: **0.116**

Thus, **product-level dynamics** explain most of the promotional sales variance. The item's inherent popularity (Fast/Medium/Slow) determines the success of a promotion more than store-level factors.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          SalesQuantity   R-squared:                       0.116
Model:                            OLS   Adj. R-squared:                  0.116
Method:                 Least Squares   F-statistic:                 1.895e+04
Date:                Fri, 18 Apr 2025   Prob (F-statistic):               0.00
Time:                        02:19:49   Log-Likelihood:            -5.5690e+06
No. Observations:             1873557   AIC:                         1.114e+07
Df Residuals:                 1873543   BIC:                         1.114e+07
Df Model:                          13
Covariance Type:            nonrobust
================================================================================================================
                                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
Intercept                                       2.2299      0.017    133.090      0.000       2.197       2.263
IsPromotion[T.True]                             0.0978      0.019      5.173      0.000       0.061       0.135
C(ProductCluster)[T.Low]                       -1.4803      0.021    -71.706      0.000      -1.521      -1.440
C(ProductCluster)[T.Max]                        3.1316      0.022    140.026      0.000       3.088       3.175
C(ProductCluster)[T.Medium]                    -0.8611      0.019    -45.320      0.000      -0.898      -0.824
C(StoreCluster)[T.Low]                         -0.5243      0.022    -23.301      0.000      -0.568      -0.480
C(StoreCluster)[T.Max]                          0.6870      0.021     33.139      0.000       0.646       0.728
C(StoreCluster)[T.Medium]                      -0.3055      0.019    -16.322      0.000      -0.342      -0.269
IsPromotion[T.True]:C(ProductCluster)[T.Low]    0.0391      0.023      1.686      0.092      -0.006       0.085
IsPromotion[T.True]:C(ProductCluster)[T.Max]    0.4216      0.025     16.666      0.000       0.372       0.471
IsPromotion[T.True]:C(ProductCluster)[T.Medium] -0.0551     0.022     -2.559      0.010      -0.097      -0.013
...
==============================================================================
```

f. Is there any significant difference betIen promotion impacts of Fast versus Slow items?

Yes.

- Fast items (baseline): uplift = ~**+0.0978**
- Slow items (IsPromotion[T.True]:C(ProductCluster)[T.Low]): **+0.0391**, *p = 0.092* (not strongly significant)

- Medium items: negative coefficient **-0.0551**, *p = 0.010*

So, Fast items benefit the most, while Slow and Medium **items** show limited or even negative reactions to promotion campaigns.

g. Is there any significant difference betIen promotion impacts of Fast versus Slow stores?

Yes.

- Fast stores (baseline): uplift ≈ **+0.0978**
- Slow stores: IsPromotion[T.True]:C(StoreCluster)[T.Low] = **-0.0549**, *p = 0.040*
- Medium stores: **-0.0291**, not statistically significant (*p = 0.193*)

This confirms that Fast stores are significantly more responsive to promotional campaigns than others, reinforcing that baseline store performance influences promotion success.

B.

B-1. Metrics

I evaluated the data with three measures:

- **Mean Absolute Error (MAE).** MAE expresses the average forecast error in the original unit of "items per day". I use it because it is easily interpretable by the business ("on average I miss by 2.5 units") and is not dominated by a few very large errors.
- **Mean Squared Error (MSE)** and its square root **RMSE**. It highlights whether the model occasionally makes very large mistakes.
- **Mean Absolute Percentage Error (MAPE).** MAPE normalises the error by actual demand and enables comparisons across products with very different sales levels.

B-2. Numerical Results

The initial model provides a **reasonable baseline**. With a **MAE of 2.55 units** and an **RMSE of 5.61**, it predicts the general direction of sales correctly but struggles with daily fluctuations and outliers. This level of accuracy is **good enough for weekly planning**, but not yet reliable for **daily operational decisions**, especially in products with irregular sales.

To improve this, I applied the **Holt-Winters method**, which better captures seasonality and trend. This significantly reduced forecast errors at the weekly level and made the model more robust for future promotions.

B-3.Main Problem Points Causing Poor Fits:

- **Simplified Clusters:**
  Using only three clusters for products and stores limits the model's ability to capture detailed sales patterns.
- **Missing Variables:**
  Key variables like pricing, inventory, and competitor activity were not included in the model, limiting its accuracy. In addition, the available product hierarchy (ProductGroup1 and

ProductGroup2) was not used. This structure could improve the model by allowing grouped behavior across similar items (e.g., all T-shirts under "Top Clothing").

- **No Time-Series Structure:**
  The regression model ignores trends and seasonality. Even Holt-Winters struggled due to irregular promotional spikes.

B-4. Planned Improvements (Including Holt-Winters)

To improve forecast performance, I propose using the following additional data:

- **Discount percentages** – to measure true promo impact.
- **Competitor promotions** – for external influence.
- **Calendar variables** – holidays, weekends, paydays.
- **Product hierarchy (ProductGroup1/2)** – to generalize behavior across similar SKUs.

To address the limitations of the initial model, I applied a **multiplicative Holt-Winters model** on weekly totals. It significantly improved accuracy by capturing seasonality and trends:

- **MAE dropped from 2.55 (daily OLS) to 0.10 (weekly HW)**
- **RMSE dropped from 5.61 to 0.13**

To further improve the model, I propose:

1. **Cluster-level Holt-Winters** – to reflect different seasonal patterns by SKU group.
2. **Feature enrichment** – include real discount rates, calendar effects, and competitor signals.
3. **Product hierarchy modeling** – use group-level patterns to stabilize low-volume SKUs.