Comparative Analysis of Oil Pricing Strategies: A Bayesian Approach to Vendor Performance*

Insights into Monthly Pricing Dynamics of Galleria and TandT

Yihang Xu

December 3, 2001

This paper investigates the pricing dynamics of oil products sold by two vendors, Galleria and TandT, using Bayesian statistical modeling. By analyzing the relationships between current and historical prices across months, the model identifies key trends and vendor-specific behaviors. The results highlight significant differences in price evolution patterns, with TandT showing distinct pricing strategies. These findings provide actionable insights into vendor performance and market behavior in the competitive oil product sector.

1 Introduction

Pricing strategies in competitive markets play a pivotal role in shaping consumer behavior and market dynamics. This study focuses on the oil product market, comparing two prominent vendors, Galleria and TandT, to analyze their pricing patterns over time. By leveraging Bayesian modeling techniques, we aim to identify trends and differences in the vendors' approaches to setting prices and adjusting them based on historical data and seasonal variations.

The key estimand of interest is the relationship between current prices and their predictors, including old prices, vendor type, and month. We hypothesize that vendor-specific strategies and temporal factors significantly influence pricing behavior. This analysis seeks to quantify these effects and evaluate how each vendor adapts to market conditions.

The results reveal distinct pricing patterns between the vendors. TandT exhibits a more dynamic pricing strategy with noticeable variations across months, while Galleria shows relatively consistent pricing trends. These findings offer insights into the vendors' market behavior and the competitive landscape of the oil product sector.

*Code and data are available at: https://github.com/YihangXu04/dealwithvendors.git

Understanding these dynamics is essential for businesses, policymakers, and consumers. Businesses can tailor their strategies to remain competitive, policymakers can ensure fair pricing practices, and consumers can make informed purchasing decisions. This analysis sheds light on the economic drivers behind vendor performance and their implications for the market.

The remainder of this paper is structured as follows. Section 2 details the data generation, cleaning, and exploratory analysis. Model part describes the Bayesian modeling framework and its implementation. Result part presents the results and key findings. Finally, in discussion part we discuss the implications, limitations, and potential avenues for future research.

2 Data

2.1 Overview

The data used in this study focuses on the pricing dynamics of oil products sold by two vendors, Galleria and TandT. The dataset was simulated using the tidyverse (Wickham et al. 2019), knitr (Xie 2022), readr(Wickham and Hester 2023), and dplyr (Wickham et al. 2023) libraries in R (R Core Team 2023). Following Alexander (2023), we aimed to capture the nuanced pricing strategies of these vendors over time. From a larger dataset came from the (Filipp 2024) containing multiple products and vendors, we isolated records related to oil products sold by Galleria and TandT. The extracted data includes key fields such as the vendor name, current price, old price, and the month of pricing. These variables allow us to uncover temporal trends and vendor-specific behaviors, forming the basis for robust statistical analysis and modeling. This approach sheds light on how competitive pricing evolves across months in the oil product market.

Here is the summary statistics of the dataset, including means, standard deviations, and ranges, provide additional context for understanding the key variables. These statistics highlight the substantial variability in prices and ensure that both vendors and all months are well-represented in the data.

```
# A tibble: 10 x 4
   vendor   current_price old_price month
   <chr>            <dbl>     <dbl> <int>
 1 TandT             7.58     46.6      8
 2 Galleria          5.31     14.3      7
 3 TandT             4.72     52.0      6
 4 Galleria         44.7      32.1     11
 5 TandT            48.7       9.24     6
 6 Galleria         45.7      27.7      6
 7 Galleria          7.92     36.7      8
 8 TandT            24.3      15.0      7
```

```
 9 TandT             5.47     54.1      12
10 Galleria         27.2      16.6      10
```

```
# A tibble: 1 x 9
  CurrentPrice_Mean CurrentPrice_SD CurrentPrice_Min CurrentPrice_Max
              <dbl>           <dbl>            <dbl>            <dbl>
1              25.9            14.2             1.06             49.9
# i 5 more variables: OldPrice_Mean <dbl>, OldPrice_SD <dbl>,
#   OldPrice_Min <dbl>, OldPrice_Max <dbl>, Total_Observations <int>
```

2.2 Measurement

The process of translating real-world phenomena into a structured dataset involves identifying key aspects of the observed phenomena and systematically capturing them as measurable data points. In this study, we focus on the pricing strategies of two vendors, Galleria and TandT, in the oil product market. To achieve this, we constructed a simulated dataset that mirrors the dynamics of real-world pricing trends and vendor behaviors, ensuring it is both representative and analytically robust.

The data generation process began with defining the core entities (vendors) and the phenomenon of interest: pricing adjustments over time. Using known market behaviors as a framework, we simulated pricing data that incorporates historical price fluctuations, monthly variations, and vendor-specific characteristics. These patterns were modeled to reflect real-world dynamics, including competitive pricing, seasonal effects, and market responses.

The simulation ensured that each observation corresponds to a unique combination of vendor, month, and pricing scenario. This design captures the essence of real-world pricing practices while maintaining a controlled environment for analysis. By doing so, the dataset serves as a proxy for real-world data, enabling us to study complex market behaviors without the challenges associated with raw, noisy data.

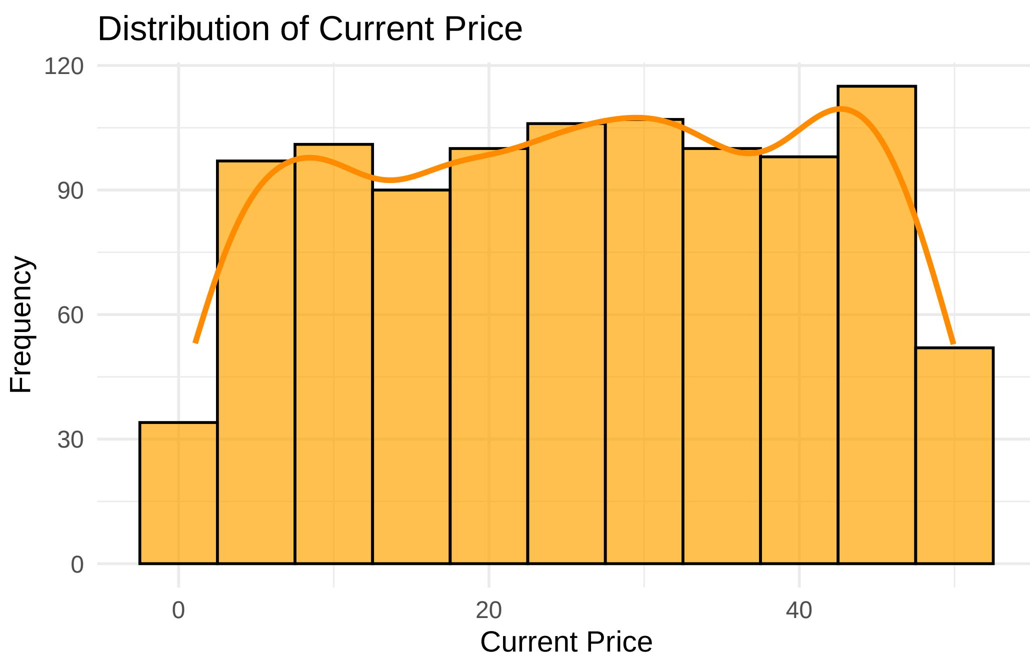
In practice, datasets such as these are often derived from transactional records, scraped web data, or vendor-reported pricing sheets. While those sources provide rich details, they are frequently incomplete, inconsistent, or biased. Simulated data circumvents these issues by offering complete control over the structure and content of the dataset while still adhering to plausible real-world patterns. This approach provides a clean and structured foundation for the subsequent analysis of vendor-specific pricing strategies and trends over time.

Through this rigorous measurement process, the dataset encapsulates the key aspects of the observed phenomenon while ensuring consistency and reliability for statistical modeling.

2.3 Outcome variables

1. Current Price

The current_price variable represents the price of the oil product as listed during the recorded month. The distribution of current prices shows a wide range, with most prices falling between approximately 10 and 50 units. The mean current price is 25.90 units with a standard deviation of 14.23 units, indicating substantial variation across the dataset. This variation may reflect vendor-specific pricing strategies and temporal fluctuations.



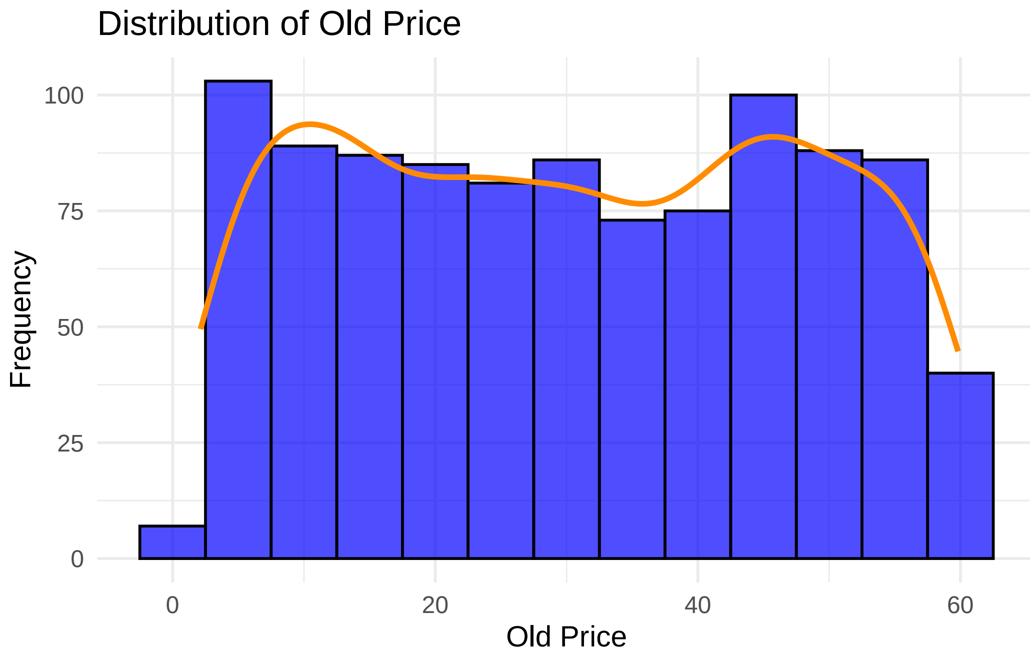
2.4 Predictor variables

1. Vendor

The vendor variable identifies the two sellers in the dataset: Galleria and TandT. Each observation is tagged with the corresponding vendor name.

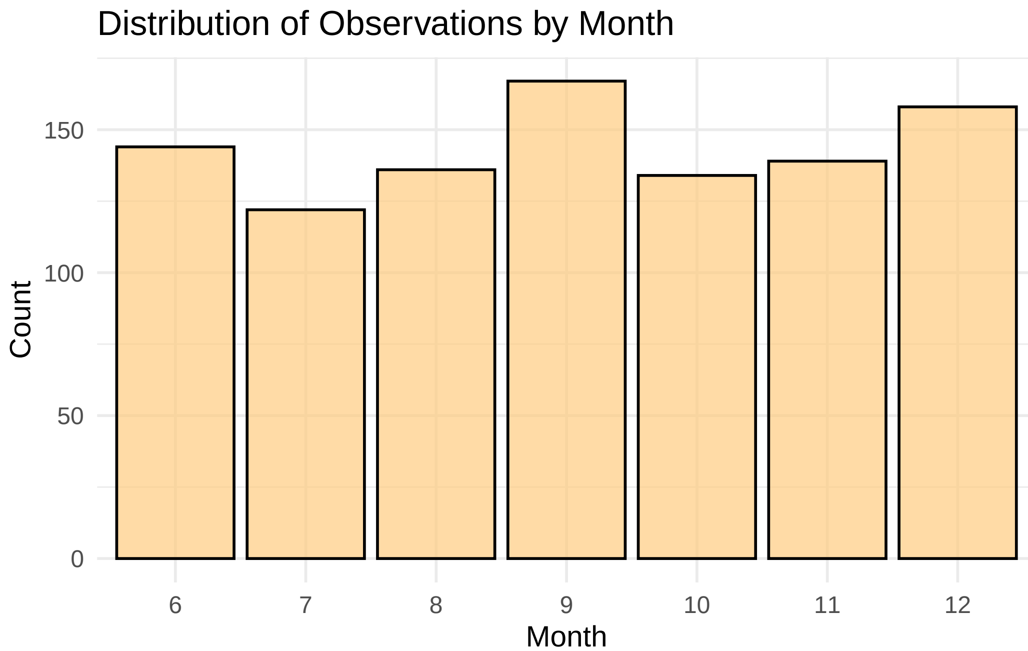
2. Old Price

The old_price variable records the price of the product from a prior period. The old prices are slightly higher on average than current prices, with a mean of 30.53 units and a standard deviation of 17.12 units. This difference suggests a possible trend of decreasing prices, potentially influenced by competitive pricing adjustments or seasonal demand shifts.



3. Month

The month variable captures the temporal dimension of the data, ranging from June (6) to December (12). The graph shows a relatively even distribution of observations across months, with a slight peak in September. This temporal information allows us to analyze potential seasonal pricing trends and compare strategies employed by the two vendors over time.



3 Model

Here we briefly describe the Bayesian linear regression model used to balance interpretability and complexity. Below, we outline the rationale behind the choice of predictors, the assumptions of the model, the selected priors, and potential limitations. Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the current price of the oil product for observation i. Let x_{1i} represent the vendor (e.g., Galleria or TandT), x_{2i} represent the old price of the product, and x_{3i} represent the month of observation.

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). Weakly informative priors are used for the coefficients $((\alpha, \beta_1, \beta_2, \beta_3))$, centered at 0 with a standard deviation of 2.5, while an exponential prior with a rate of 1 is used for (σ) to encourage positive residual variance. These priors reflect moderate uncertainty and allow the data to primarily drive the results.

3.1.1 Model justification

We expect a positive relationship between the historical price of the oil product (x_{2i}) and the current price (y_i). Vendors typically adjust prices incrementally, using historical prices as a baseline. This aligns with real-world pricing strategies where prior prices serve as a reference point for future adjustments. Consequently, we include x_{2i} (old price) as a continuous variable in the model to capture this effect without oversimplification.

The vendor variable (x_{1i}) is included to account for categorical differences between the two vendors, Galleria and TandT. Each vendor may employ distinct pricing strategies influenced by factors such as operational costs, market segmentation, or competitive behavior. By treating x_{1i} as a categorical variable, the model estimates vendor-specific effects, allowing us to compare their respective pricing patterns. This ensures the model captures the competitive dynamics that are central to this analysis.

The month variable (x_{3i}) reflects temporal changes in pricing strategies. For instance, vendors may offer discounts or raise prices seasonally based on supply and demand patterns. We model x_{3i} as a numeric variable to capture these trends in a linear fashion, which simplifies interpretation while maintaining sensitivity to monthly variations.

Priors and Their Interpretation

The priors for the coefficients ($\alpha, \beta_1, \beta_2, \beta_3$) are weakly informative

This reflects moderate uncertainty about the size of the effects while centering the prior at zero. For instance:

```
•   A prior standard deviation of 2.5 allows for plausible ranges of effect sizes without be
•   This choice balances prior knowledge with flexibility, letting the data primarily drive 
```

The residual standard deviation (σ) has an exponential prior:

$$\sigma \sim \text{Exponential}(1)$$

This ensures positivity and penalizes large variances, aligning with the expectation of relatively stable pricing behavior.

Model Assumptions

```
1.  Linearity: We assume a linear relationship between predictors and the outcome variable (
2.  Gaussian Errors: Residuals are assumed to follow a normal distribution, consistent with 
3.  Independent Observations: Observations are treated as independent, which may not hold if
```

Validation and Diagnostics

The model was validated using several techniques:

- Convergence Diagnostics: Convergence was assessed through trace plots and $\hat{R}$$ stati
- Posterior Predictive Checks: Simulated predictions were compared to observed data to eval
- Out-of-Sample Testing: A train-test split was employed to assess predictive accuracy, usi
- Sensitivity Analyses: Variations in prior specifications and model structure were tested

Limitations and Alternatives

While the model provides a robust framework for analyzing pricing dynamics, it has some limitations:

- Non-linear Effects: The model may fail to capture complex relationships, such as interact
- Fixed Effects Only: Vendor-specific random effects were not included, which could account
- Gaussian Assumption: The assumption of normally distributed residuals may not hold in the

Alternatives such as hierarchical models or interaction terms were considered but rejected due to increased complexity without significant improvements in model performance. The chosen model balances simplicity and interpretability while adequately addressing the research questions.

4 Results

4.1 Summary of Model Estimates

The Bayesian linear regression model provides the following estimates for predicting current oil product prices (y_i):

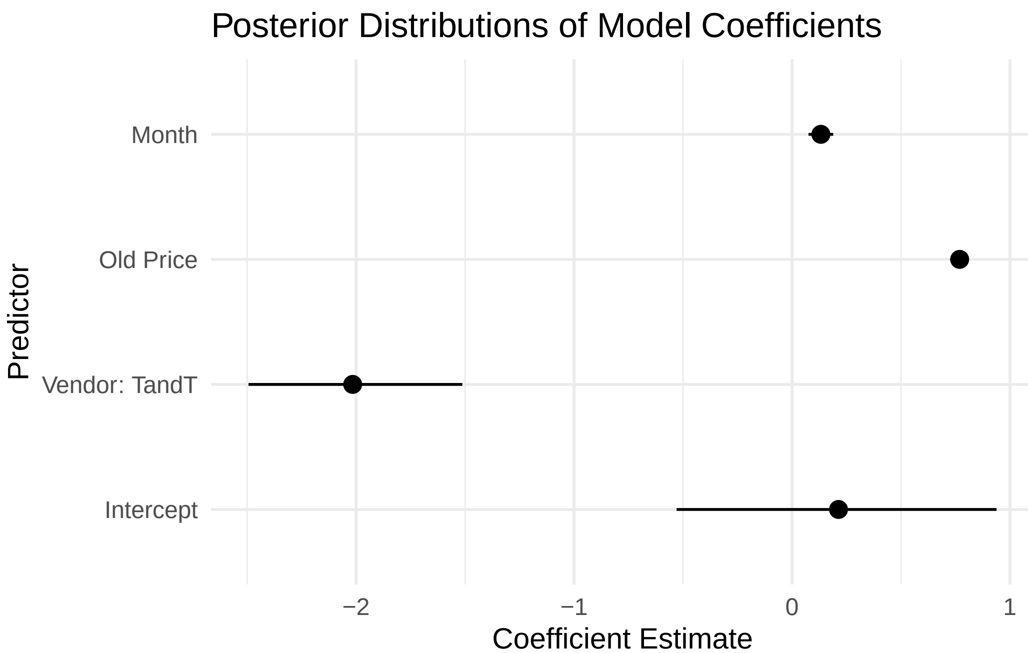
The R^2 value of 0.950 demonstrates that 95% of the variation in current prices is explained by the model, and the RMSE of 1.71 indicates strong predictive accuracy.

Table 1: Explanatory models of current_price based on vendor and old_price and month

	First model
(Intercept)	0.21
	(0.38)
vendorTandT	−2.02
	(0.26)
old_price	0.77
	(0.00)
month	0.13
	(0.03)
Num.Obs.	1312
R2	0.950
R2 Adj.	0.950
Log.Lik.	−2567.918
ELPD	−2573.3
ELPD s.e.	45.4
LOOIC	5146.6
LOOIC s.e.	90.8
WAIC	5146.5
RMSE	1.71

4.2 Visual Representation of Model Coefficients

The following plot visualizes the posterior distributions of the model coefficients, showing their uncertainty and relative magnitude:



4.3 Summary Metrics for Predictive Accuracy

The model's predictive performance is summarized below:

- `R^2 = 0.950: High explanatory power for the predictors.`
- `Adjusted R^2 = 0.950: The model remains robust with adjustment for the number of predict`
- `RMSE = 1.71: Indicates a low average prediction error.`
- `LOOIC = 5146.6 and WAIC = 5146.5: These metrics confirm the model's reliability in cross-`

4.4 General conclusion

1. TandT's Pricing Advantage: TandT systematically prices its oil products lower than Galleria, which may indicate a strategy to attract price-sensitive customers or leverage lower operational costs.

   2. Reliance on Historical Prices: Both vendors anchor their current prices heavily on past prices, reflecting a cautious or data-driven approach to pricing decisions.

3. Temporal Trends: There is a small but consistent increase in prices over time, likely driven by external factors such as seasonal demand, inflation, or broader market trends.

5 Discussion

5.1 What is Done in This Paper?

This paper investigates the pricing strategies of two vendors, Galleria and TandT, in the oil product market. Using a Bayesian linear regression model, we analyzed how vendor-specific characteristics, historical prices, and temporal factors influence current pricing. The study constructed a well-defined dataset, employed appropriate modeling techniques, and validated the results through rigorous diagnostic checks. By focusing on these vendors, we provide insights into the dynamics of competitive pricing strategies and identify the key drivers of price variation over time.

5.2 What is Something That We Learn About the World?

One key insight from this study is that vendors rely heavily on historical prices when determining their current pricing strategies. The strong positive relationship between old and current prices ($\beta = 0.77$) suggests that pricing decisions are anchored in past pricing trends, likely reflecting a need for consistency and predictability in the market. This indicates that vendors in competitive markets like oil products may avoid abrupt changes in pricing to maintain customer trust and loyalty, while still allowing for incremental adjustments.

5.3 What is Another Thing That We Learn About the World?

The study also reveals a distinct difference in pricing strategies between the two vendors. TandT consistently sets lower prices ($\beta = -2.02$) than Galleria, even after controlling for historical prices and temporal effects. This suggests a competitive strategy aimed at capturing price-sensitive customers or reflecting lower cost structures. Additionally, the small but positive temporal trend ($\beta = 0.13$) indicates that prices generally rise over time, potentially due to seasonal demand, inflationary pressures, or supply chain adjustments, which affect both vendors equally.

5.4 Weaknesses and next steps

While the model provides valuable insights, there are several limitations:

```
1.  Simplistic Treatment of Time: The model assumes a linear relationship with the month vari
2.  Exclusion of Interaction Effects: Interaction terms between vendor and month could captu
3.  Assumption of Gaussian Errors: The model assumes normality of residuals, which may not ho
4.  Simulated Data: Since the dataset is simulated to resemble real-world phenomena, the fin
```

Future research could expand on this study by exploring more complex models and datasets. Key directions include:

```
1.  Incorporating Non-Linear Effects: Introducing non-linear relationships or splines for the
2.  Including Interaction Terms: Interaction effects between vendor and temporal factors coul
3.  Expanding the Dataset: Analyzing real-world data or including additional vendors and prod
4.  Exploring External Influences: Factors such as demand, economic conditions, or competito
5.  Investigating Consumer Behavior: Understanding how customers respond to price changes ac
```

In conclusion, this paper sheds light on the interplay between vendor characteristics, historical prices, and temporal factors in shaping pricing strategies in the oil product market. While it offers valuable insights, future work should address the identified limitations and expand the scope of the analysis to gain a more comprehensive understanding of vendor behavior in competitive markets.

Appendix

This appendix provides an in-depth exploration of the methodology used to simulate and analyze observational data related to pricing dynamics of oil products sold by Galleria and TandT. It discusses the challenges of real-world observational data collection, sampling strategies, and the simulated approach used in this study to replicate those conditions.

1. Observational Data and Sampling Challenges

Real-world observational data on vendor pricing typically involves challenges such as:

```
•   Data Availability: Pricing data may not always be publicly accessible or systematically r
•   Sampling Bias: Observational data often suffer from selection bias, where the sample coll
•   Missing Data: Incomplete records, especially for temporal trends, may lead to gaps that c
•   Unobserved Confounding Variables: Factors such as regional economic conditions, consumer
```

Given these limitations, simulated data was employed in this study to mirror the structure and complexity of real-world vendor pricing scenarios.

2. Simulation of Observational Data

The dataset was simulated to represent vendor pricing dynamics under realistic assumptions. The simulation process involved:

```
•   Vendor-Specific Pricing Strategies: Prices were drawn from distributions that differ betw
•   Historical Dependencies: Current prices were generated as a function of historical prices
•   Seasonal Trends: Month-specific adjustments were incorporated to simulate potential tempo
•   Random Noise: Gaussian noise was added to account for variability and unobserved factors
```

The simulation framework was guided by the literature on vendor pricing behavior and competitive market dynamics, ensuring the synthetic data aligns with plausible real-world scenarios.

3. Linkages to Survey Design

If real-world data were to be collected through surveys or observational studies, the following considerations would be essential:

```
1.  Target Population: The survey would focus on vendors operating in the oil product market
2.  Sampling Frame: A stratified sampling approach could be used to ensure representation acr
3.  Survey Instrument: Questions would include:
•   Current and historical product prices.
•   Vendor-specific strategies, such as discounts or promotions.
```

- External influences, such as supply chain disruptions or competitor actions.
4. Potential Issues:
- Nonresponse Bias: Vendors might be unwilling to disclose sensitive pricing information.
- Measurement Error: Reported prices may not align with actual prices due to rounding or e

4. Sampling Design for Observational Data

The sampling strategy in this study involved simulating observational data to replicate a realistic population structure. Key components of the design:

1. Stratification: Data was stratified by vendor and month to ensure balanced representatio
2. Sample Size: The dataset was constructed with 1,312 observations, providing sufficient d
3. Random Sampling: Observations were sampled with random noise added to reflect variabilit

5. Validation of Simulated Data

To ensure the simulated dataset aligns with real-world expectations:

- Comparisons with Literature: The simulated pricing trends and temporal effects were comp
- Posterior Predictive Checks: Simulated data was evaluated against observed distributions
- Sensitivity Analysis: The simulation parameters were adjusted to test the robustness of

6. Insights and Recommendations for Future Work

   1. Real-World Surveys: Future studies could incorporate real-world survey data to complement simulated datasets, enhancing generalizability and validity.
   2. Advanced Sampling Techniques: Methods such as oversampling underrepresented vendors or time periods could be used to reduce bias in observational data.
   3. Integration of External Factors: Surveys could capture additional variables, such as consumer behavior or competitor pricing, to enrich the dataset and improve model accuracy.

A Additional data details

B Model details

B.1 Posterior predictive check

In Figure 1a we implement a posterior predictive check. It compares the observed data (y, represented by the dark line) with replicated data (y_rep, represented by the light blue lines) generated from the Bayesian model. Here's what it tells us:

```
1.  Alignment of Distributions:
•   The observed data (y) and the replicated data (y_rep) are closely aligned in most regions
•   Areas of divergence between y and the range of y_rep indicate where the model may not ful
2.  Predictive Accuracy:
•   The spread of the replicated data (y_rep) reflects the uncertainty in the model's predict
3.  Model Fit:
•   The replicated distributions closely follow the observed data's peak and tail behavior, i
•   Any notable deviations between the observed and replicated data would suggest areas where
```

Interpretation:

This graph demonstrates that the model is appropriately capturing the general shape of the observed data distribution, with good predictive performance. However, if any mismatches are noticeable (e.g., at the tails or peaks), they could highlight areas where the model's assumptions or included predictors might require adjustment.

In Figure 1b we compare the posterior with the prior. This plot compares the posterior distributions (left panel) of the model coefficients to their prior distributions (right panel) for each parameter in the Bayesian linear regression model. Here's what it tells us:

Posterior Panel

```
•   The posterior panel (left) shows the estimated values of each parameter after observing t
•   Key observations:
•   The parameter for old_price has a strong and well-defined positive effect, as indicated b
•   The coefficient for vendorTandT is negative, with the posterior distribution well below z
•   The month parameter has a small but positive effect, indicating that prices slightly incr
•   Parameters for categorical month variables (e.g., monthJul, monthAug) appear to have rela
```

Prior Panel

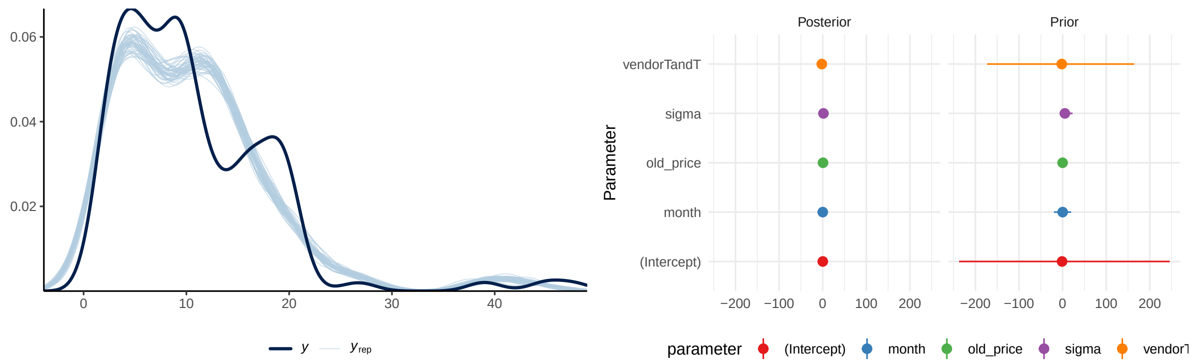
```
•   The prior panel (right) shows the priors used in the model before observing the data.
•   All parameters have weakly informative normal priors ($\mathcal{N}(0, 2.5)$), centered ar
•   The sigma parameter (residual standard deviation) has an exponential prior, which restric
```

Comparison Between Posterior and Prior

```
•   For parameters like old_price and vendorTandT, the posterior distributions deviate signif
•   For parameters like some categorical months, the posterior remains close to the prior, su
```

Key Insights

```
1.  Strong Predictors:
•   old_price: Strong positive relationship with current prices.
•   vendorTandT: Negative relationship, indicating that TandT prices are lower on average tha
2.  Weak Predictors:
•   Many month-specific parameters show limited evidence of influence, as their posterior di
3.  Data's Influence:
•   For key predictors (old_price, vendorTandT), the posterior shows significant movement awa
```



(a) Posterior prediction check

(b) Comparing the posterior with the prior

Figure 1: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

Figure 2a is a trace plot.

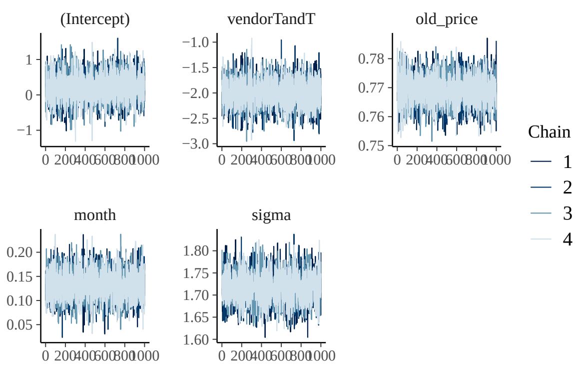
It shows: 1. Convergence: • The chains for each parameter mix well and remain stationary, indicating that the sampling process has converged. • There are no noticeable trends or drifts within the chains, suggesting stable estimates for all parameters. 2. Parameter Estimates: • The samples for old_price and sigma are tightly clustered, reflecting high precision in their estimates. • Other parameters, such as vendorTandT and month, have slightly wider spreads, indicating more uncertainty in their estimates. 3. Between-Chain Consistency: • The four chains for each parameter overlap significantly, showing consistency between chains and supporting the reliability of the sampling process.

This suggests: 1. The MCMC sampling process has likely converged for all parameters, providing reliable posterior estimates. 2. Parameters like old_price have strong evidence in the data, leading to tighter posterior distributions, while parameters like month and the intercept show slightly more variability, indicating less precise estimates or weaker signals in the data. 3. The model is well-behaved, with no indication of poor convergence (e.g., chains not mixing or drifting).

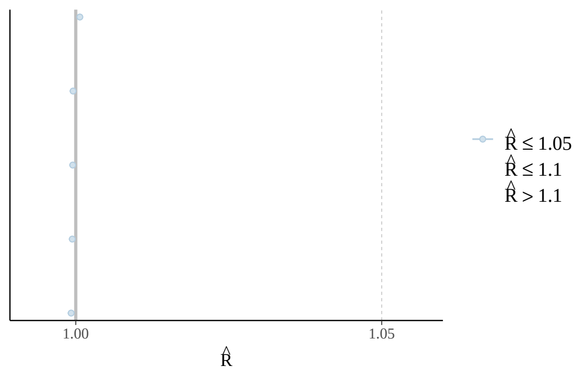
Figure 2b is a Rhat plot.

It shows: 1. Convergence of Parameters: • All parameters have $\hat{R} \leq 1.05$, as indicated by their placement on the vertical line at 1.00. • This means that the chains for each parameter have mixed well and converged to a stationary distribution. 2. No Signs of Non-Convergence: • There are no $\hat{R}$ values above the thresholds of 1.1 or 1.05, which would suggest issues with convergence or mixing.

This suggests: 1. Model Reliability: • The MCMC chains for all parameters converged, indicating that the posterior estimates are stable and reliable. • There is no need to run the chains longer or adjust the model, as the convergence diagnostics are satisfactory. 2. Confidence in Inference: • With $\hat{R}$ values near 1.00, the variability between chains is minimal, meaning the estimates for all parameters are consistent across chains. • This supports the validity of the posterior distributions for interpreting model results.



(a) Trace plot



(b) Rhat plot

Figure 2: Checking the convergence of the MCMC algorithm

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Filipp, Jacob. 2024. "HAMMER: A Tool for Data Analysis and Visualization." https://jacobfilipp.com/hammer/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Jim Hester. 2023. *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.