Main Insights and Recommendations

Short Summary of Key Findings

1. Price Variation by Product

1. **Gasoline**
   - **Generally commands the highest average selling price** among the three primary products (Gasoline, Diesel, Ethanol).
   - The distribution often centers around **4 BRL**, but can extend beyond **8 BRL** in some outlier cases—possibly driven by regional taxes, specialized branding, or unique local market structures.
2. **Diesel**
   - Often priced just below Gasoline on average, with a **median in the low- to mid-3 BRL** range.
   - Despite a slightly lower baseline, Diesel can also see **extreme outliers** above 8 BRL, suggesting localized conditions (e.g., remote industrial hubs, supply chain bottlenecks).
3. **Ethanol**
   - Tends to be cheaper overall, typically **2–3 BRL** per liter in median ranges.
   - Even with lower average prices, **outliers** can appear, possibly reflecting periods of ethanol shortage or demand spikes in certain markets.

Implication:
Each product follows a different pricing dynamic and volatility profile, underscoring the necessity of **product-specific pricing and procurement strategies**. Monitoring local supply, tax structures, and brand positioning is particularly critical for Gasoline and Diesel, where price variability is greater.

2. Purchase Price Dynamics

- **Most purchase prices** lie between **1 BRL and 2.5 BRL**, with fewer extreme outliers compared to selling prices.
- The **right-skewed tail** of purchase prices occasionally exceeds 4 BRL, aligning with situations where supply constraints, transportation costs, or niche suppliers drive up costs.

Potential Margin Impacts:
- **Consistently higher avg_preco_venda (sale price) vs. avg_preco_compra (purchase price)** supports a stable markup for many transactions.
- **Outlier scenarios** with very high or low purchase prices can compress or inflate margins in specific localities or time periods, indicating the need for **close supplier management** and possible renegotiation of contracts.

3. Margin Analysis & Outliers

1. **Typical Margins**:
   - The **boxplots** often show **10–30%** as a common margin range (corresponding to ~0.3–0.7 BRL over purchase).
2. **Negative or Extremely High Margins**:
   - Some transactions yield **negative spreads**, hinting at promotions, local price wars, or potential data entry errors.
   - **High-margin outliers** (above 50% or more) could be tied to remote or captive markets, premium branding, or short-term supply disruptions.
3. **Product-Specific Margin Profiles**:
   - Gasoline and Diesel typically exhibit **greater margin volatility**, whereas Ethanol shows somewhat more stable but still significant swings.

Implication:
Investigating extreme margin values (both positive and negative) can uncover hidden costs, inefficiencies, or strategic opportunities. Scrutinizing these transactions can reveal areas for **improved operational controls** or **competitive advantage** in targeted markets.

4. Geographic & Market Concentration Insights

4.1. State-Level Concentration (HHI Scores)

- **High HHI States (e.g., MA, MT, AM)**:
  - Highly concentrated; fewer dominant players; higher potential for **price premiums** but also **regulatory scrutiny**.
- **Lower HHI States (e.g., DF, RS)**:
  - More competitive; typically **lower average prices**, narrower margins due to **multiple strong competitors**.

4.2. Number of Establishments (n_estabelecimentos)

- **Highly Skewed**:
  - Most observations under 50 establishments, but some municipalities exceed 1,000.
  - Large urban centers (or aggregated reporting) can distort averages, so treating these areas separately in models is wise.

4.3. Geographic Price Variation & Correlation

- **Remote vs. Urban**:
  - **Remote or rural municipalities** often face higher transportation costs, leading to **elevated selling prices** and (often) higher margins.

  - **Urban centers** show somewhat lower median prices (due to greater competition) but can have **very high outliers** (brand premiums, overhead costs).
- **Correlation Declines Over Distance**:
  - Nearby cities often track each other's prices, whereas **far-flung municipalities** can diverge significantly due to different supply routes, local taxes, and demand factors.

Implication:
Region-specific factors—**logistics, market concentration, establishment density, and local taxation**—all play key roles in shaping price and margin. A **"one-size-fits-all"** approach to pricing will miss many nuances; **robust regional segmentation** is essential.

5. Brand Coverage & Overlapping Networks

1. **Universal State Presence**:
   - Major brands (e.g., White Label, Blue Label, Green Label, Purple Label) show coverage in all states, though not necessarily with the same market share.
2. **Competitive Pressure**:
   - Because multiple brands co-exist in most states, local competition is common. This underscores the value of **localized brand strategies**—especially where a brand's presence or reputation varies.
3. **Potential for Supply Chain Optimization**:
   - If a corporate entity owns multiple brands, there may be **transport synergies** or shared distribution that can cut costs.

   - Independently owned brands may pursue **third-party logistics contracts** or distribution partnerships to improve coverage and inventory efficiency.

Implication:
Network-level decisions—such as re-routing shipments, consolidating brands, or optimizing distribution points—can drive **significant cost savings** and help maintain price competitiveness. Overlapping footprints require careful management to **avoid cannibalization** (if brands are under one umbrella) or to **target competitor vulnerabilities** (if operating independently).

6. Time-Series Trends & Volatility

1. **Long-Term Upward Trend**
   - From **2004 to mid-2010s**, prices rose gradually, reflecting inflationary pressures and global oil market fluctuations.
2. **Post-2016 Acceleration**
   - Steeper climbs suggest **currency issues, macroeconomic shifts,**

and possibly new policy changes—impacting all three major fuels.

3. **Rising Volatility Since ~2018**
   - **12-month rolling standard deviation** doubled or tripled, approaching ~1.0 BRL, indicating larger and more frequent price swings.
   - High volatility amplifies **profit risk** and requires **agile pricing strategies** (e.g., daily or weekly adjustments, automated dynamic pricing).

Cluster Analysis of historical prices identifies **distinct phases** (e.g., lower-price era vs. high-volatility era). Each phase or cluster calls for **different inventory, hedging, and pricing tactics**.

7. Actionable Recommendations

1. **Pricing & Margin Management**
   - **Dynamic Pricing**: In high-volatility contexts (post-2018), automate or frequently review prices.
   - **Margin Outlier Investigation**: Identify negative or excessively high margins; rectify data errors or exploit market opportunities.
2. **Regional Segmentation**
   - **Focus on High-HHI States**: With fewer competitors, consider cautious price increments but watch for regulatory attention.
   - **Low-Margin, Competitive Zones**: Improve operational efficiency, possibly partner for supply discounts or explore brand differentiation.
3. **Supply Chain & Logistics**
   - **Optimize Distribution**: Re-route shipments to reduce transportation costs in remote markets.
   - **Inventory Buffering**: In volatile price environments, holding strategic reserves could mitigate cost spikes, but watch carrying costs.
4. **Brand Strategy & Network Coordination**
   - **Position Brands by Region**: Use premium branding where local demand supports higher prices; deploy cost-competitive labels in price-sensitive or highly competitive areas.
   - **Prevent Cannibalization**: If multiple brands fall under one corporate owner, define **clear market segments** for each label.
5. **Time-Based Forecasting & Scenario Planning**
   - Segment historical data into **price/volatility regimes** (clusters) to refine forecasts.
   - Develop **scenario-based business plans**: e.g., best case (steady

market), moderate (gradual price growth), worst case (high volatility & supply disruptions).

6. **Regulatory & Policy Monitoring**
    - Track ongoing **tax or subsidy changes**, environmental regulations, and import/export policy shifts that can abruptly impact price or supply.

    - Engage in **lobbying** or stakeholder dialogues where localized tax structures severely skew costs.

Modeling for forecasting

- **Model type**: RandomForestRegressor

- **Test set RMSE**: 1.0275

- **Test set R^2**: -0.2517

Financial Impact
- **Total Actual Revenue**: $4,441,103.35
- **Revenue Prediction Error**: $725,492.66 *(16.34%)*
- **Total Actual Profit**: $493,765.82
- **Profit Prediction Error**: $80,475.89 *(16.30%)*

1. Model Performance & Forecast Quality

Balancing Optimism and Caution

- **Conservative Forecasts**: The model tends to **underestimate** actual values. This can serve as a **built-in safety margin**, preventing overestimation of demand or revenue—helpful if budgets or inventory are tight.

- **Risk of Missed Upside**: Underestimating ~75% of the time suggests we might leave money on the table in strong markets. Addressing this bias could unlock **additional profit opportunities**.

Interpreting the Metrics

- **Negative R^2**: At **-0.2517**, the model underperforms a simple average baseline on this particular test set. This underscores the **complexity of the data** and signals significant room to improve feature engineering or hyperparameter tuning.

- **RMSE of 1.0275**: In a revenue/profit context, this level of error—though not trivial—may still be **actionable** for broad strategic decisions.

- **Prediction Errors (~16%)**: While the absolute deviations are notable, they provide a **starting point** for further optimization. The model can be incrementally refined to reduce these gaps.

2. Financial Impact Analysis

Revenue & Profit

- **Conservative Baseline**:
  - Predicted revenue ($3.71M) vs. actual ($4.44M).
  - Predicted profit ($413K) vs. actual ($494K).
- Even though these forecasts trail real-world values by roughly **16%**, this shortfall can be **corrected** with ongoing model improvements (e.g., adding relevant external data or revisiting feature selection).

Practical Takeaways

- **Managing Downside Risk**: With an underestimate, the business is less likely to overcommit resources.
- **Opportunity Costs**: In a bullish scenario (high market demand), undershooting forecasts may result in **stockouts** or **inadequate staffing**, limiting revenue capture.

3. Business Implications

1. **Underestimation Bias**
   - **Safe Inventory Approach**: Because the model undershoots, you're less likely to end up with large surpluses.
   - **Growth Potential**: If demand spikes, you risk missing revenue because capacity or investment might be set too low.
2. **Profitability Considerations**
   - **Conservative Projections** can help secure financing or manage costs under stable conditions.
   - **High-Volatility Markets**: Rapid price changes or demand surges may amplify the risk of leaving profits on the table.
3. **Strategic Planning**
   - **Risk-Averse Budgeting**: Forecasts at 16% below actual can serve as a **"worst-case" or baseline** scenario.
   - **Flexibility Needed**: Supplement the model's predictions with real-time market signals or rapid re-forecasting methods to **capture upswings** more effectively.

4. Recommendations & Next Steps

1. **Refine the Model**
   - **Feature Expansion**: Incorporate macroeconomic variables, competitor moves, seasonal indices, and marketing spend to capture demand fluctuations more accurately.

   - **Hyperparameter Tuning & Segmentation**: Explore separate models for different product lines, regions, or customer segments where the relationship between inputs and revenue/profit may differ.
2. **Continuous Monitoring**
   - **Rolling Retraining**: Regularly update the model with fresh data, ensuring it learns from recent market shifts.

   - **Real-Time Dashboards**: Track actual vs. predicted performance and alert decision-makers when deviations exceed a critical threshold (e.g., $\pm 10\%$).
3. **Scenario & Contingency Planning**
   - **Multiple Forecast Scenarios**: Use the model's conservative outputs for baseline planning, but also create optimistic/"best-case" scenarios to inform opportunistic decisions.

   - **Buffer Stocks & Optionality**: In markets prone to rapid demand surges, maintain a small **inventory cushion**—countering the model's underestimation bias.
4. **Operational Safeguards**
   - **Threshold Tweaks**: If you use profit thresholds for go/no-go decisions, consider adjusting them upward to **compensate** for the model's conservatism.

   - **Sensitivity Analyses**: Periodically test how changes in certain inputs (price, demand, cost) might yield different outputs—and plan your supply chain or marketing campaigns accordingly.

In Summary

Despite the model's current tendency to **underpredict** revenue and profit by around 16%, these forecasts provide a **solid conservative baseline**. From a **risk management** perspective, underestimation can shield you from **overcommitting resources**, but it also risks **missing out on potential gains** in a vibrant market.

By iterating on feature sets, refining hyperparameters, and monitoring real-time performance, you can **gradually close the gap** between predicted and actual outcomes—striking a **better balance** between **risk aversion** and **growth potential**. The key is to **leverage the model as a starting point** for budgetary discipline while **complementing** its forecasts with **flexible, data-**

driven adjustments to seize emerging opportunities.

Possible next Steps

1. Model Improvements

1. **Try Additional Algorithms**: Use **AutoML** (e.g., H2O.ai, auto-sklearn) to quickly evaluate new models (LightGBM, CatBoost) and tune hyperparameters more effectively.

2. **Advanced Approaches**: Test **Bayesian models** or **Prophet** for time-series behavior.

3. **Feature Selection**: Apply **RFE** or SHAP to prioritize the most impactful features.

4. **Expand Hyperparameter Search**: Move beyond simple grids; consider Bayesian optimization for better results.

2. Experiment Tracking & Deployment

1. **MLflow**: Track experiments, metrics, and parameters for reproducibility.

2. **Dockerize & Deploy**: Package the model in a Docker container for consistent, scalable production use.

3. **CI/CD**: Automate testing, building, and deployment of the Docker image.

3. Risk Section

1. **Underestimation Bias**
   - **Risk**: Missing out on revenue during demand spikes.
   - **Mitigation**: Include real-time indicators and retrain often.
2. **Data Gaps & Quality**
   - **Risk**: Key features may be absent or incomplete.
   - **Mitigation**: Expand data sources; ensure continuous quality checks.
3. **Overfitting**
   - **Risk**: Model performs poorly with big market shifts.
   - **Mitigation**: Rolling retrains, scenario testing.
4. **Computational Complexity**
   - **Risk**: Long runtimes with extensive hyperparameter searches.
   - **Mitigation**: Use cloud resources, set cost/time limits.
5. **Interpretability vs. Accuracy**

- **Risk**: Complex models are harder to explain.
- **Mitigation**: Use explainable AI tools (SHAP, LIME).