541 Data Mining Final Report
Frequent Pattern Mining

By
Jittapatana (Patrick) Prayoonpruk
&
Nexaly Orellana

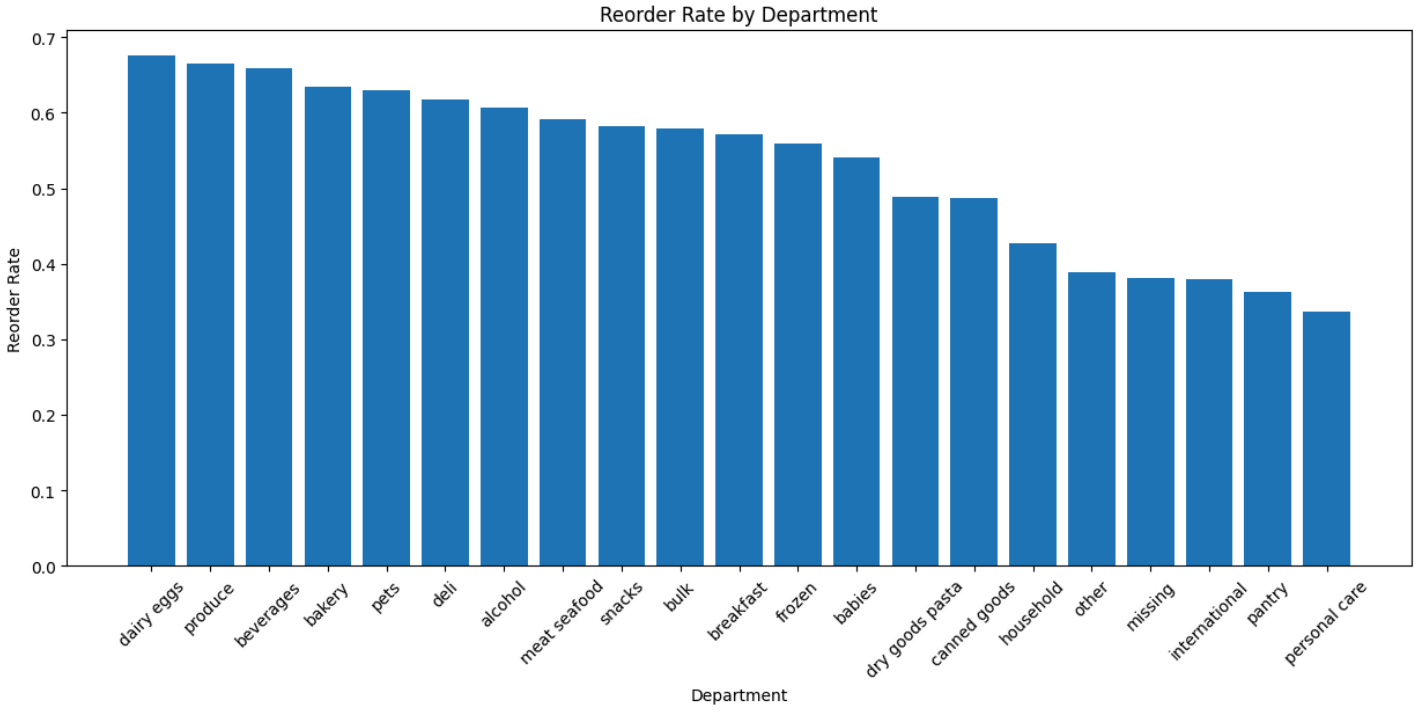
INTRODUCTION

This report was compiled using data from the Instacart Market Basket Analysis dataset that can be found on Kaggle.com. This data represents order information from over 200,000 Instacart users, and compiles order information for over 3 million grocery orders made by those users. For every user represented in the data, there is expected to be anywhere between four and one hundred order data points to analyze. These data points contain information about the products purchased in each order, the week and hour of day the order was placed, and the time elapsed between orders. Also included in the Instacart data is information about the different grocery store departments and the aisles associated with those departments. Using this data, we have analyzed five important pieces of statistical information, found frequent pattern associations, and used machine learning techniques to anticipate order size and product demand.

Statistical Analysis

To analyze the Instacart Market Basket Analysis we agreed on five pieces of statistical information which we deemed to be important, these would be:

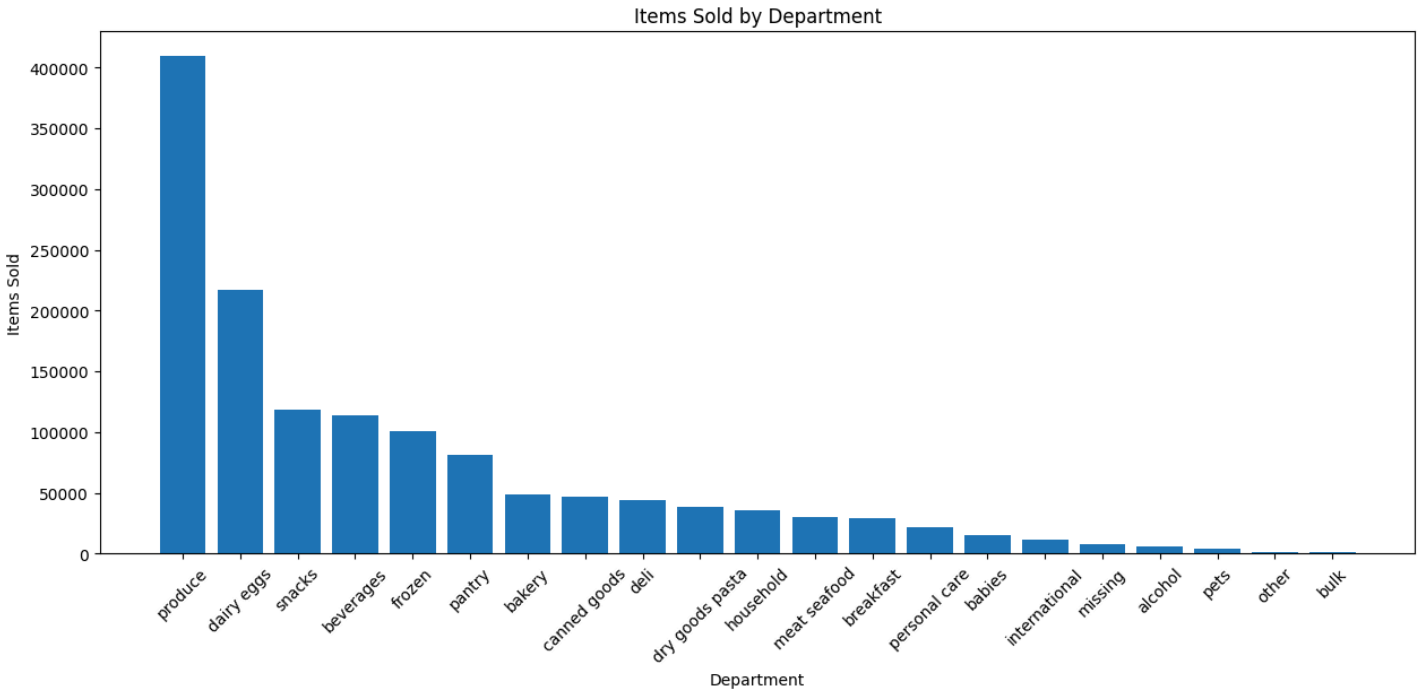
1. Reorder rate by department



From this bar graph we can see that the “dairy eggs” department has the highest reorder rate just below 70%. Similarly we can also see that personal care products have the lowest reorder rate at just over 30%. By analyzing the graph we can see that departments that sell consumables such as “dairy eggs”, “produce”, and “beverages” are reordered more often than compared to departments that might sell

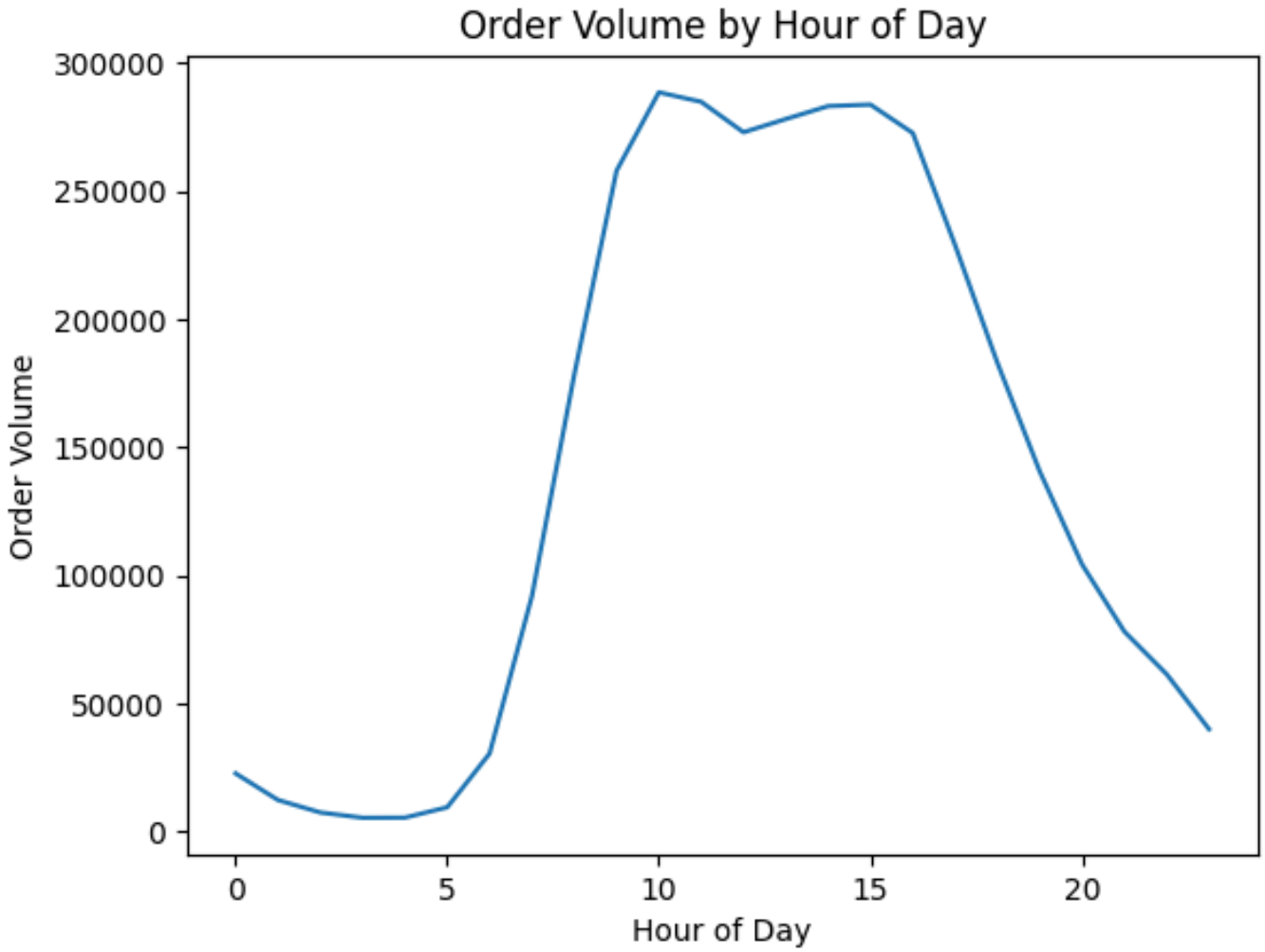
non-consumables, like “personal care”, “pantry”, and “household” items. This information can be useful when stores or instacart create promotions, as adding a coupon or promotion to an item bought in a low reorder department such as personal care, might boost sales for that department. Additionally this information is useful for inventory management, since departments that have higher reorder rates need to be restocked more than departments that have lower order rates.

2. Items sold by department



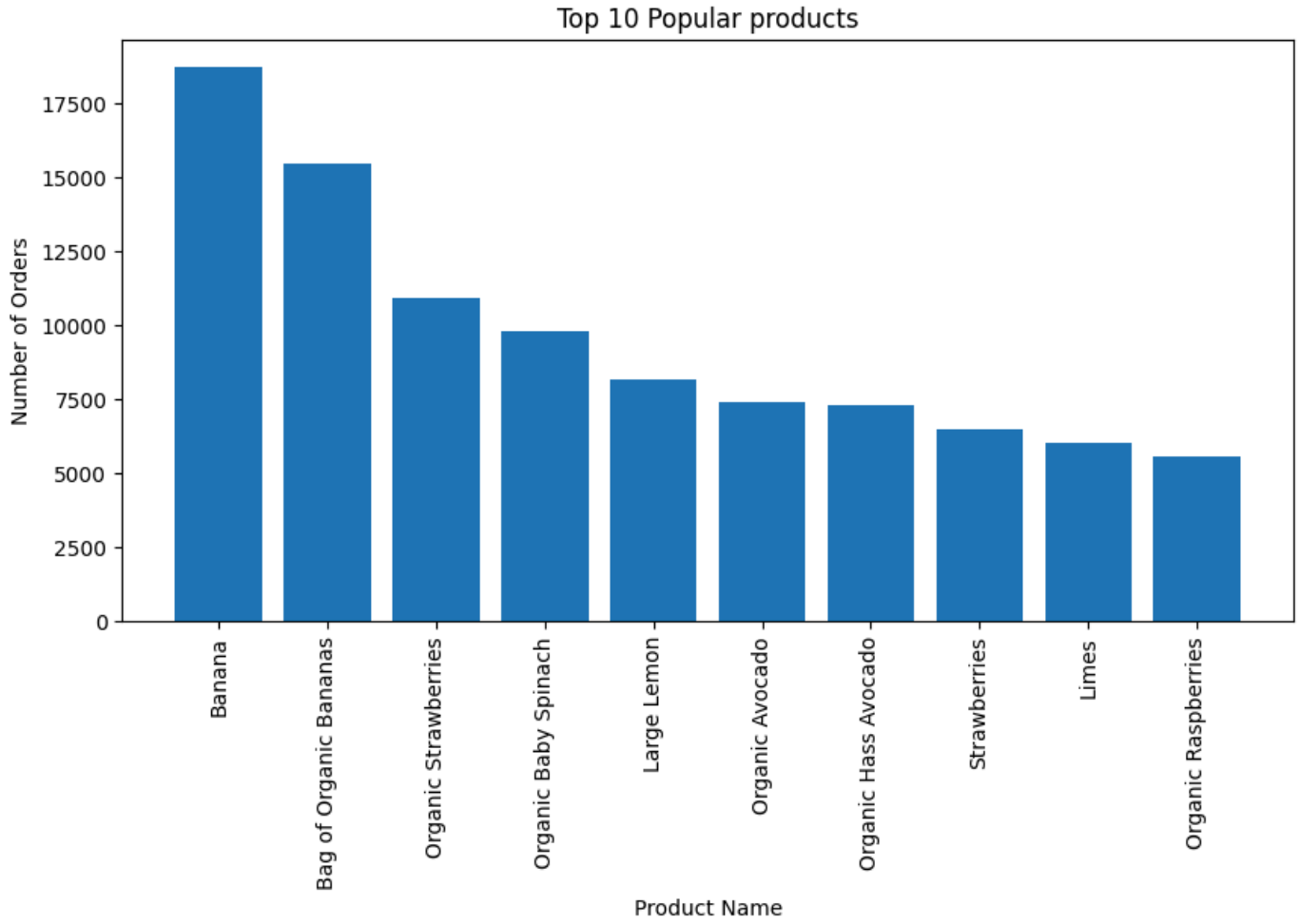
From this bar graph it is shown that produce has the largest amount of items sold, at over 400,000 units, indicating that produce items are a popular order item that is frequently purchased. On the other end of the spectrum however, we see that the bulk department has the lowest number of items sold with under 1500 items sold for the department. This lower level of units sold tells us that items in the bulk department are relatively unpopular and are not frequently purchased. Similar to our previous analysis, this bar graph helps us solidify the idea that consumable products are more popular than non-consumable products. This information can be helpful when analyzing consumer preferences which influence buying trends.

3. Order volume by hour of day



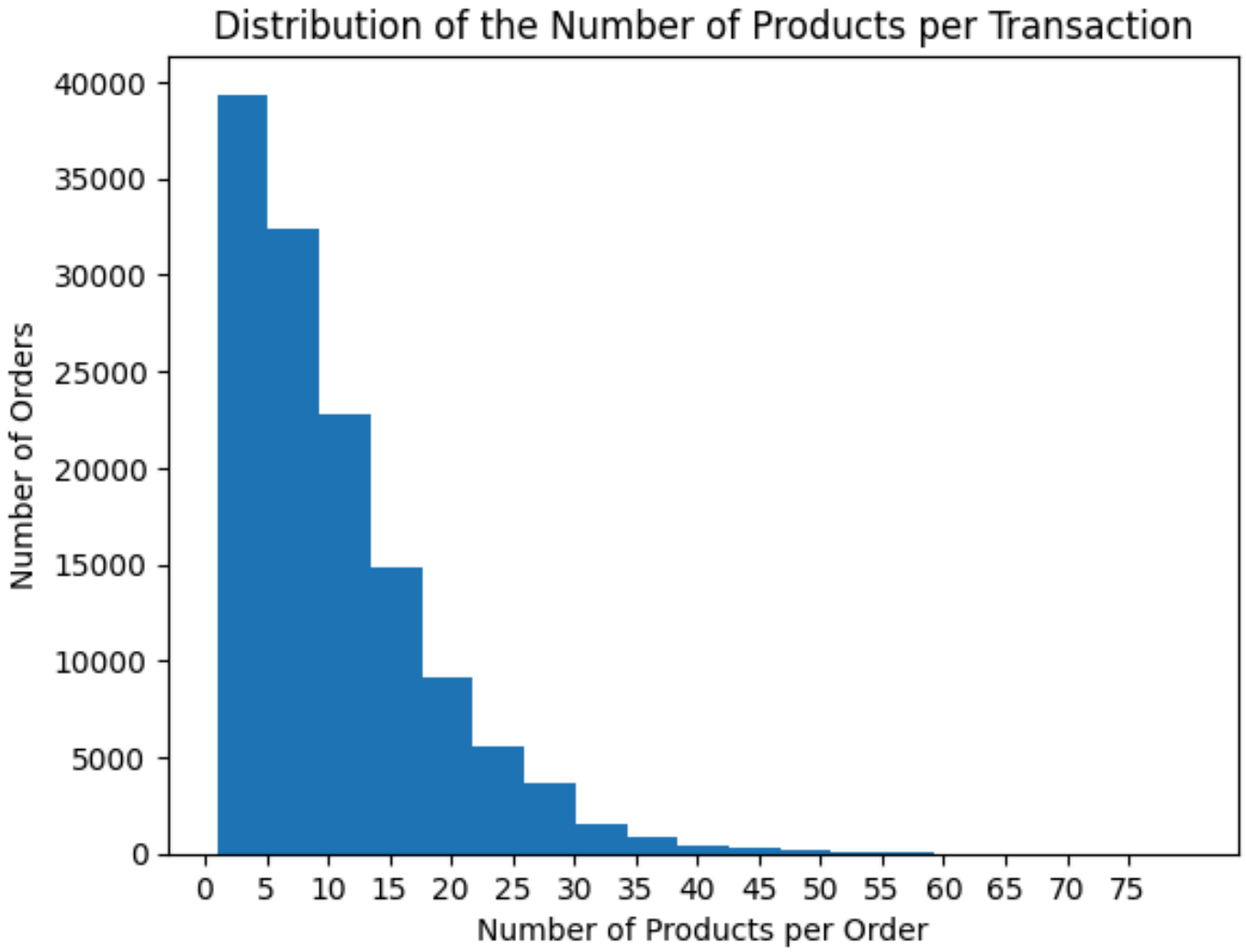
This line graph represents the amount of orders placed throughout any given day. We can see that there are a relatively low amount of orders made between the hours of midnight and 5:00 AM. Between the hours of 5:00 AM and 9:00 AM the amount of orders increases to its peak of just under 300,000 orders. Between the hours of 9:00 AM and 4:00 PM orders seem to stay relatively close to the peak, and after 4:00 PM the amount of orders starts to decrease until midnight. This tells us that most orders are made between the hours of 10:00 AM to 4:00 PM. This type of information can help with staffing and instacart payouts. If a grocery store knows there are a higher amount of orders placed between the hours of 9:00 AM to 4:00 PM then they can staff more employees to help with the influx of instacart shoppers that might come in at those hours. Additionally, since those are peak hours with lots of orders that need to be fulfilled, Instacart can offer higher pay during those hours to incentivize drivers to complete orders during the busy period.

4. Top 10 most popular products ordered



From this bar graph we can see that bananas are the most popular product bought, being purchased in more than 17500 orders. We can also see that organic raspberries, while still popular, are in the final spot of the top 10 only being purchased in around 5500 orders. By looking at this graph we notice that all items in the top 10 are produce items, meaning that customers generally order healthy items. Additionally since all items in the top 10 are consumable items, this data adds more credibility to the hypothesis that consumable products are more popular than non-consumable products. This type of data can help with inventory management, letting ordering managers know what items are in high demand and need to be restocked on a frequent basis.

5. Distribution of the number of products per transaction



This bar graph represents the amount of products purchased throughout orders. From this we can determine that in almost 40000 orders between 1 and 5 items were purchased, we can also see that the amount of orders made decrease and the amount of items increase. We can conclude from this, that most customers are placing orders with a relatively small amount of items, and very rarely are orders with large quantities of items placed. We can see that only around 5000 orders are made requesting more than 20 items. This information can be helpful for marketing, allowing for promotions and discounts for buying in bulk or for buying items that are commonly bought together, this can help increase the amount of items bought per order, increasing revenue.

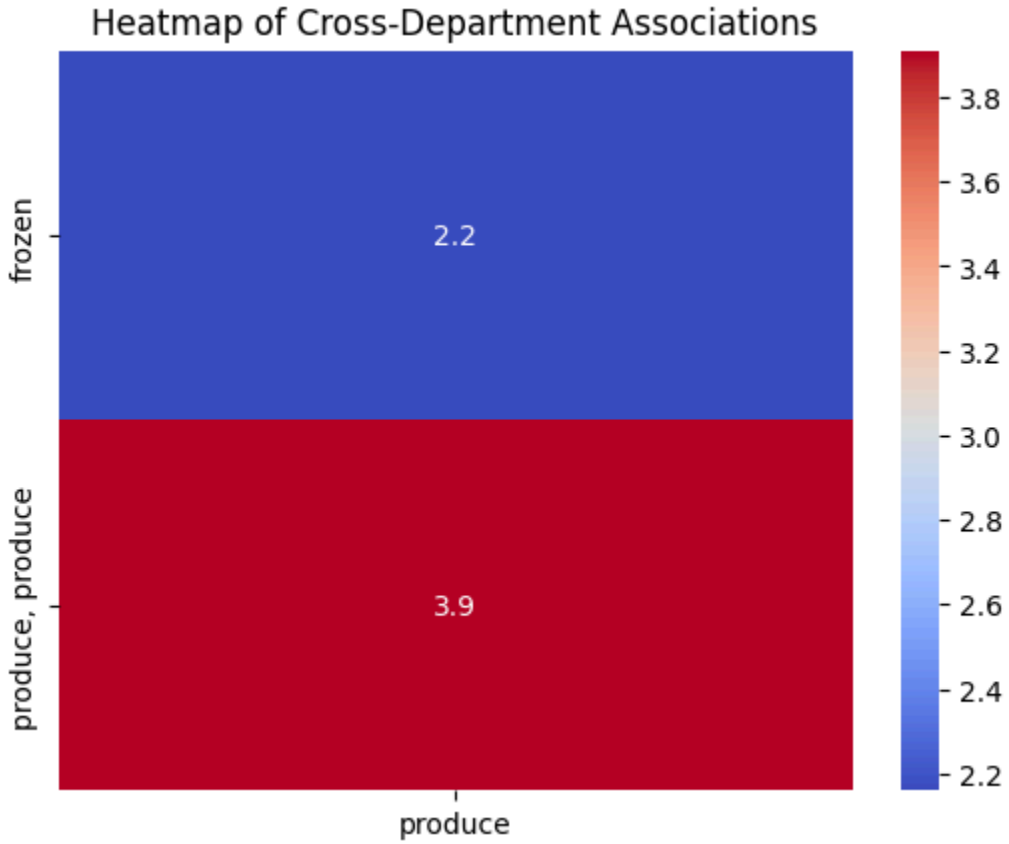
Frequent Pattern Association

To find these associations we used:

- Minimum Support = 0.005
    - Definition
        - Support measures how frequently an item or itemset appears in the dataset, calculated as the proportion of transactions containing the itemset
    - Clarification:
        - A low threshold ensures inclusion of moderately frequent itemsets without overwhelming the model with too many patterns
        - 0.005 means that the itemset must appear in at least 0.5% of the transactions, balancing between rare and common patterns in a large dataset
        - Due to the volume of transactions, larger datasets benefit from slightly lower support values
- Confidence > 0.3
    - Definition
        - Confidence measures the probability that a consequent (item B) appears in a transaction given the presence of an antecedent (item A).
    - Clarification
        - A threshold of 30% ensures a meaningful relationship between itemsets; if item A is present, item B appears with at least 30% probability
        - It avoids low-confidence rules that might not be practically useful or actionable
        - For market basket data, this value strikes a balance between high-quality rules and diversity in recommendations
- Lift > 1.5
    - Definition
        - Lift measures the strength of an association rule, comparing the observed frequency of a rule to what would be expected if the items were independent
    - Clarification
        - Lift > 1.5 ensures that the items in the rule are positively correlated and not just appearing together by chance

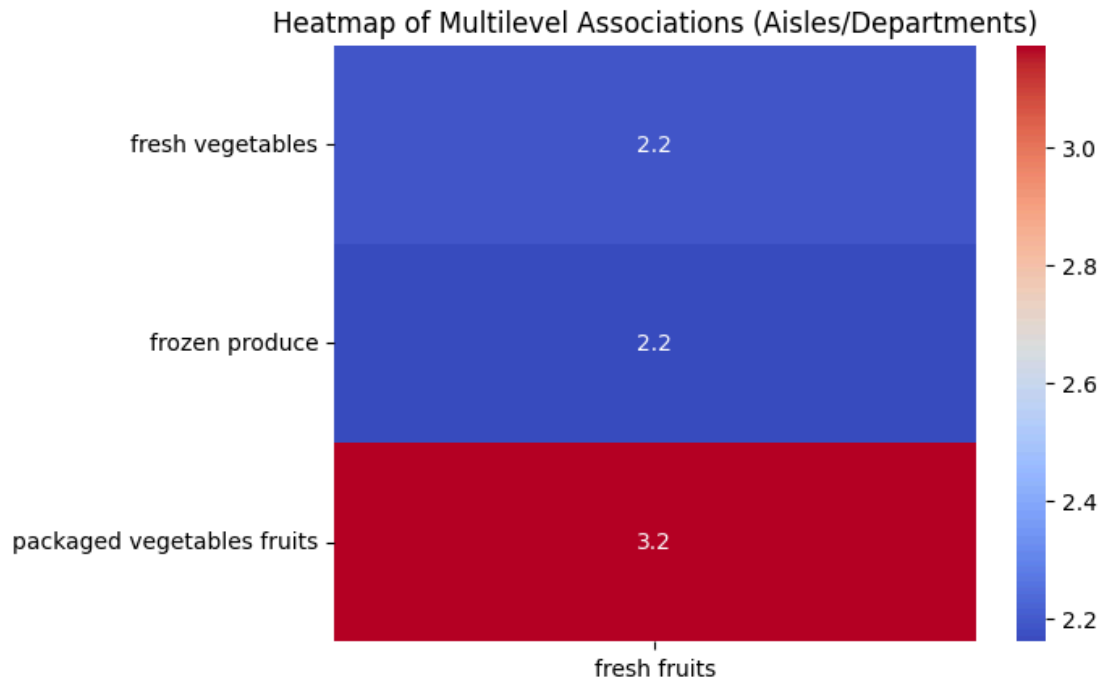
- A value above 1.5 indicates that the antecedent significantly increases the likelihood of the consequent, making the rule actionable for marketing or inventory strategies.
- Choosing 1.5 instead of a higher value allows for a broader but still meaningful set of association rules

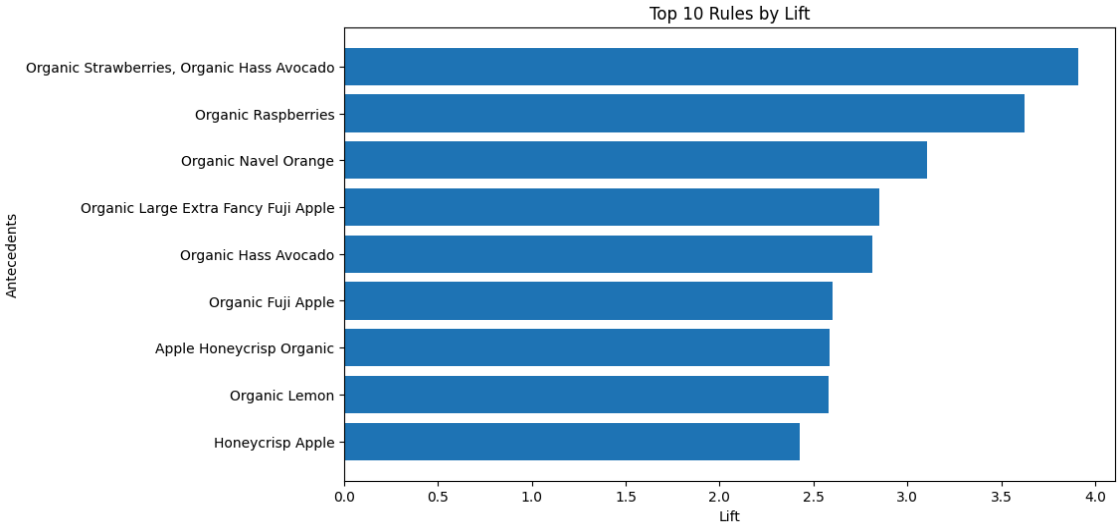
1. Strong Association Across Departments



This heatmap represents the strong associations between the produce department and other departments. From this we can deduce that produce is often ordered with other produce, as it has a lift of 3.9. Similarly, we can say that produce is also often bought with items from the frozen department, as it has a lift of 2.2. These are the only departments that have a strong association that follows our rules of having a lift greater than 1.5. This type of information can be used to design grocery layouts, allowing for departments with strong associations to be placed near each other for optimal shopping experience. This information can also be used to create cross-department promotions.

2. Multilevel Association Across Aisles in a Department





This heatmap represents the multilevel association between the fresh fruits aisle in the produce department and other aisles in the produce department. We can see from this that fresh fruit is commonly bought with packaged vegetables and fruits, as it has a high lift of 3.2. We can also see that fresh fruits are also commonly bought with items from the frozen produce aisle as well as items from the fresh vegetables aisle, as both have a lift of 2.2. The Top 10 Rules by lift graph are the rules that inform our heatmap. Like with the strong association across departments, this information can also be useful when designing optimal grocery store layouts and in creating cross-department promotions.

Extra Credit

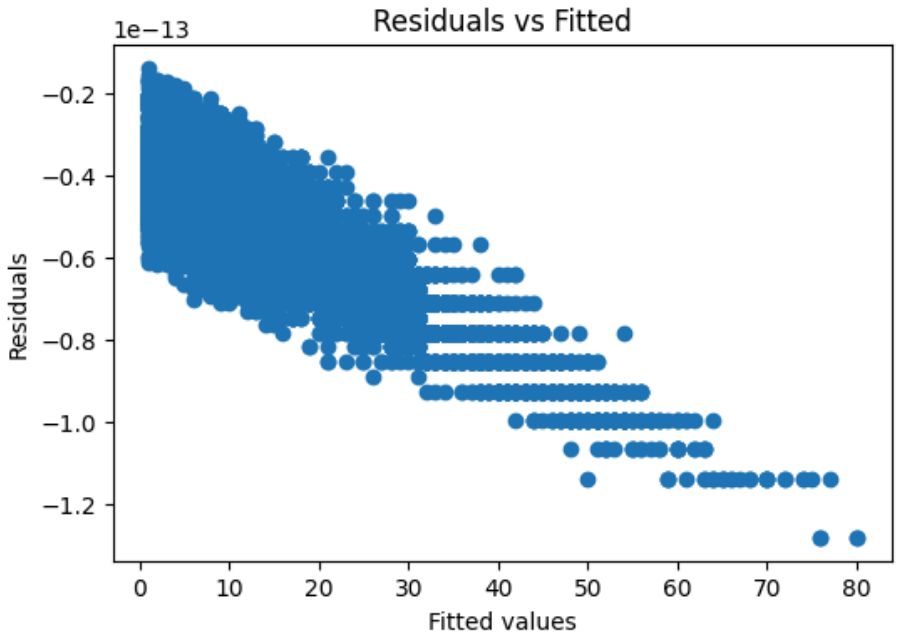
1. Predicting Order Size Using Multiple Regression

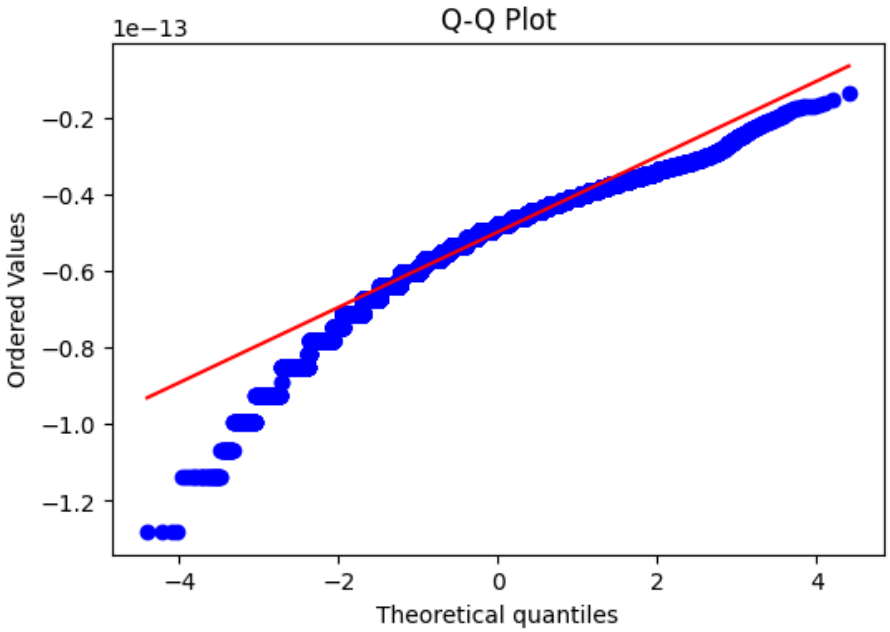
OLS Regression Results

Dep. Variable:	order_size	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	9.984e+31
Date:	Fri, 29 Nov 2024	Prob (F-statistic):	0.00
Time:	16:23:25	Log-Likelihood:	3.8302e+06
No. Observations:	131209	AIC:	-7.660e+06
Df Residuals:	131176	BIC:	-7.660e+06
Df Model:	32		
Covariance Type:	nonrobust		

The linear regression model was developed to predict order size using various predictors related to order characteristics, user behavior, day of the week, and time of the day. The dependent variable in the model was order_size, and the independent variables included order_number, days_since_prior_order, user_avg_order_size, dummy variables for days of the week (order_dow_1 to order_dow_6), and dummy variables for order hours (order_hour_of_day_1 to order_hour_of_day_23). The model summary revealed that all predictors used with significant coefficients, e.g., user_avg_order_size and order_hour_of_day_10 had significant impacts on order_size ($p < 0.05$).

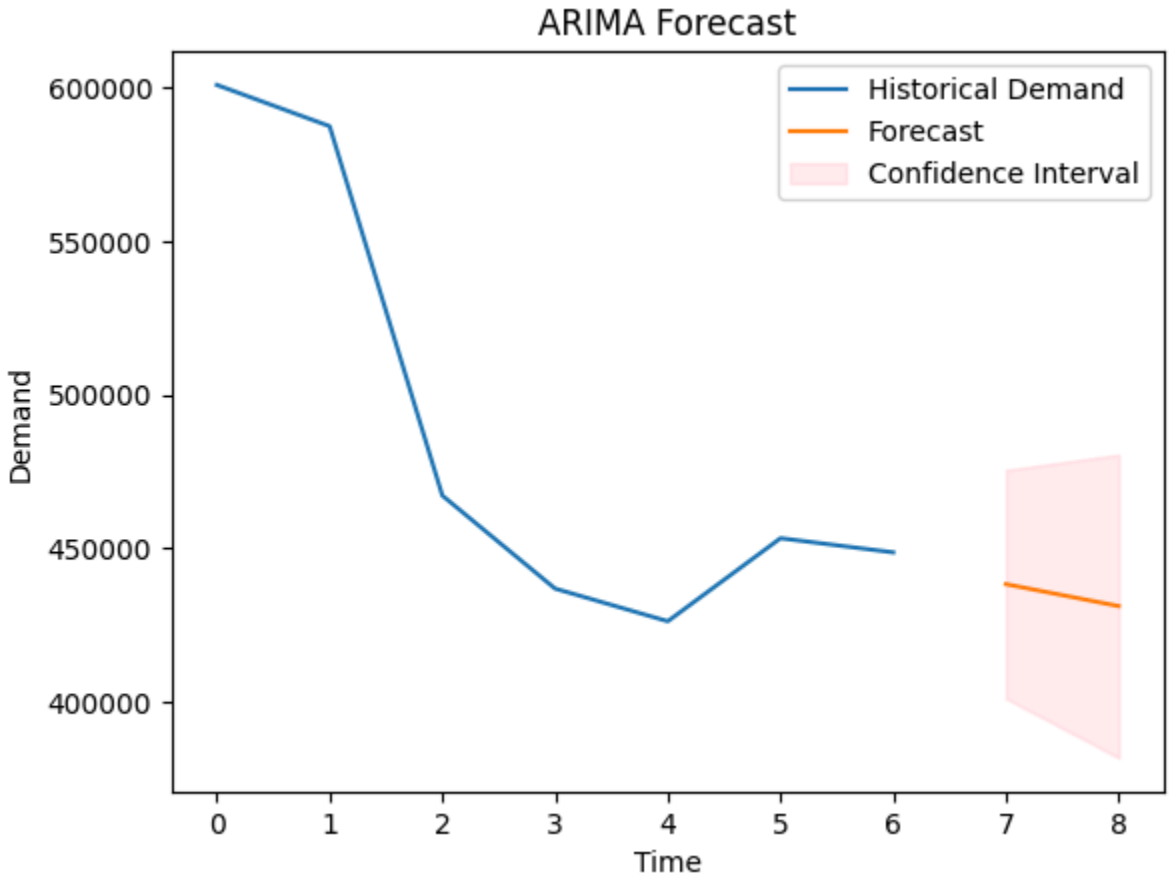
While the linear regression model provides valuable insights into the factors influencing order size, several diagnostic concerns were observed, indicating potential limitations in the model's reliability. The residual diagnostic plots showed discernible patterns, suggesting that the residuals are not randomly distributed, which violates the assumption of independence. Additionally, the QQ plot revealed deviations from normality, indicating that the residuals do not follow a normal distribution. Furthermore, evidence of heteroscedasticity (non-constant variance of residuals) was detected, which could undermine the validity of statistical inferences drawn from the model. Notably, the model achieved an R2 of 1, which strongly suggests potential overfitting, likely due to the limited size of the training data.





These observations highlight that the relationship between the predictors and the response variable may not be purely linear in nature. The patterns in the residuals and other diagnostic issues suggest that a more flexible model, such as polynomial regression, generalized additive models (GAM), or machine learning techniques like random forests or gradient boosting, might better capture the underlying relationships in the data. Future analysis should also consider increasing the sample size and exploring non-linear relationships to improve model robustness and predictive accuracy.

2. Forecasting Product Demand Over Time



An ARIMA model was applied to forecast product demand using historical data aggregated by the day of the week. The dataset was split into training and test sets, with the last two data points used for evaluation. The model's performance was assessed using the Mean Absolute Percentage Error (MAPE), which was calculated to be 3.60%. This indicates that the model's predictions were relatively accurate, with an average deviation of only 3.60% from the actual demand values in the test set.
Observations and Recommendations

The low MAPE value demonstrates the model's ability to capture demand trends effectively. However, incorporating additional features or exploring alternative models could help further enhance the accuracy and reliability of the forecasts. This analysis highlights the potential of ARIMA modeling in supporting demand planning and operational decisions.

Observations and Recommendations

Model Accuracy: The ARIMA model performed well in capturing the overall trend, as evidenced by the low MAPE and RMSE values. However, there is room for improvement in terms of precision, particularly for capturing short-term fluctuations when more training data is available.

Uncertainty Consideration: The widening confidence intervals highlight the increased uncertainty in forecasts as the model moves beyond the observed data range. This underscores the importance of regular model updates with recent data to maintain prediction accuracy.

Potential Improvements: The linear ARIMA model might have limitations in capturing potential non-linear patterns in the data. Incorporating exogenous variables, such as promotions, holidays, or other external factors, could improve the model's predictive power.

Future Steps: To further refine demand forecasting, exploring alternative models such as seasonal ARIMA (SARIMA) or machine learning approaches (e.g., Gradient Boosting or LSTM) might capture additional complexities in the data.