Instacart Customer Segmentation Analysis

Executive Summary

Instacart needed to move beyond generic marketing campaigns and implement targeted strategies based on actual customer behavior. Through analysis of over 32 million transactions across 200,000+ customers, I developed a data-driven segmentation framework that identified four distinct customer profiles and uncovered critical insights about ordering patterns, loyalty behaviors, and product preferences.

Key Findings:

- Family households (with dependents) generate 72% of all orders, with Small Families leading at 15.3M orders
- Peak ordering occurs Saturday-Sunday between 9 AM-4 PM, creating clear windows for off-peak ad scheduling
- Loyal customers order 55% more frequently than new customers, indicating strong retention value
- Produce and Dairy/Eggs departments dominate across all customer segments, accounting for the highest transaction volumes
- Regional differences had minimal impact on ordering behavior, suggesting consistent national demand patterns

Impact: Delivered actionable segmentation framework and six strategic recommendations enabling Instacart's marketing team to implement targeted campaigns aligned with customer behavior patterns rather than assumptions.

Project Overview

Client: Instacart (Online Grocery Platform)
Role: Data Analyst
Duration: 10-week project
Tools: Python (pandas, NumPy, matplotlib, seaborn), Jupyter Notebooks, Excel
Dataset Scale: 32.4M transactions, 206K customers, 50K products

The Challenge

Instacart's marketing and sales teams needed to move beyond one-size-fits-all campaigns. With a diverse customer base and varying purchasing behaviors, they sought data-driven insights to implement targeted marketing strategies that would improve campaign effectiveness and customer engagement.

Key Questions:

- When should ads be scheduled to avoid peak order times?
- Which customer segments show the highest engagement and loyalty?
- What products drive the most revenue across different customer profiles?
- How do demographics influence ordering patterns?

My Approach

Phase 1: Data Preparation & Quality Assurance

Working with datasets of this scale presented immediate challenges. The raw data arrived in four separate files requiring integration and extensive cleaning before analysis could begin.

Datasets Combined:

- **Orders:** 3,421,083 orders with temporal and user identifiers
- **Products:** 49,688 unique products with pricing and department information
- **Order-Product History:** 32,434,489 individual line items (prior orders only)
- **Customer Demographics:** 206,209 users with age, income, family status, and regional data

Data Quality Challenges & Solutions

Challenge 1: Mixed-Type Columns
The products dataset contained price columns with inconsistent data types—some stored as strings, others as floats—causing merge failures and calculation errors.

Solution: Implemented systematic type checking and conversion logic:

```
# Identified mixed types using df.dtypes and value inspection

# Converted all price fields to consistent float64 format

# Validated conversions with spot checks on random samples
```

Challenge 2: Missing Values
Discovered 206,209 missing values in the days_since_prior_order field, representing first-time orders with no prior purchase history.

Solution: Rather than impute or delete, I flagged these as "New Customer" orders and retained them for segmentation analysis. This preserved critical information about customer acquisition patterns while maintaining data integrity.

Challenge 3: Duplicate Records
Found duplicate product entries that would inflate order counts and skew department analysis.

Solution: Implemented deduplication checks:

- Identified duplicates using product_id as key
- Verified duplicates weren't valid repeated purchases
- Removed 5 duplicate product records
- Documented exclusions for audit trail

Challenge 4: Dataset Scale & Performance
The merged dataset exceeded 32M rows, causing memory issues and slow processing times in standard workflows.

Solution:

- Exported intermediate results using pickle format (faster I/O than CSV)
- Implemented chunked processing for aggregations
- Created targeted subsets for specific analyses rather than loading full dataset repeatedly
- Reduced final working dataset to essential columns only

Merge Validation

Combining four datasets required careful validation to ensure no data loss:

Merge 1: orders + orders_products_prior

- Result: 32,434,489 records (100% match)

- Validation: Both-sided merge confirmed complete join

Merge 2: Combined + products

- Result: 32,404,859 records retained

- Excluded: 29,630 records with missing product info (likely discontinued items)

Merge 3: Final + customers

- Result: 32,404,859 records

- Validation: All orders successfully matched to customer profiles

Phase 2: Customer Segmentation Strategy

Phase 2: Customer Segmentation Strategy

Rather than relying solely on traditional demographic cuts, I developed a behavioral segmentation approach that combined family status with purchasing patterns. This method better captures actual needs and usage patterns.

Primary Segmentation Framework:

I created four customer profiles based on dependents and age:

1. **Small Family** (1-2 dependents)
    - 15.3M orders (47% of total volume)
    - Core customer base
    - High-frequency users
2. **Big Family** (3+ dependents)
    - 7.6M orders (23% of total volume)
    - Second-largest segment
    - Bulk purchasing behavior
3. **Older Adult, No Kids** (Age 50+, 0 dependents)
    - 5.0M orders (15% of total volume)
    - Steady, predictable patterns
4. **Younger Adult, No Kids** (Age 18-49, 0 dependents)
    - 2.6M orders (8% of total volume)

- Smallest segment but growth opportunity

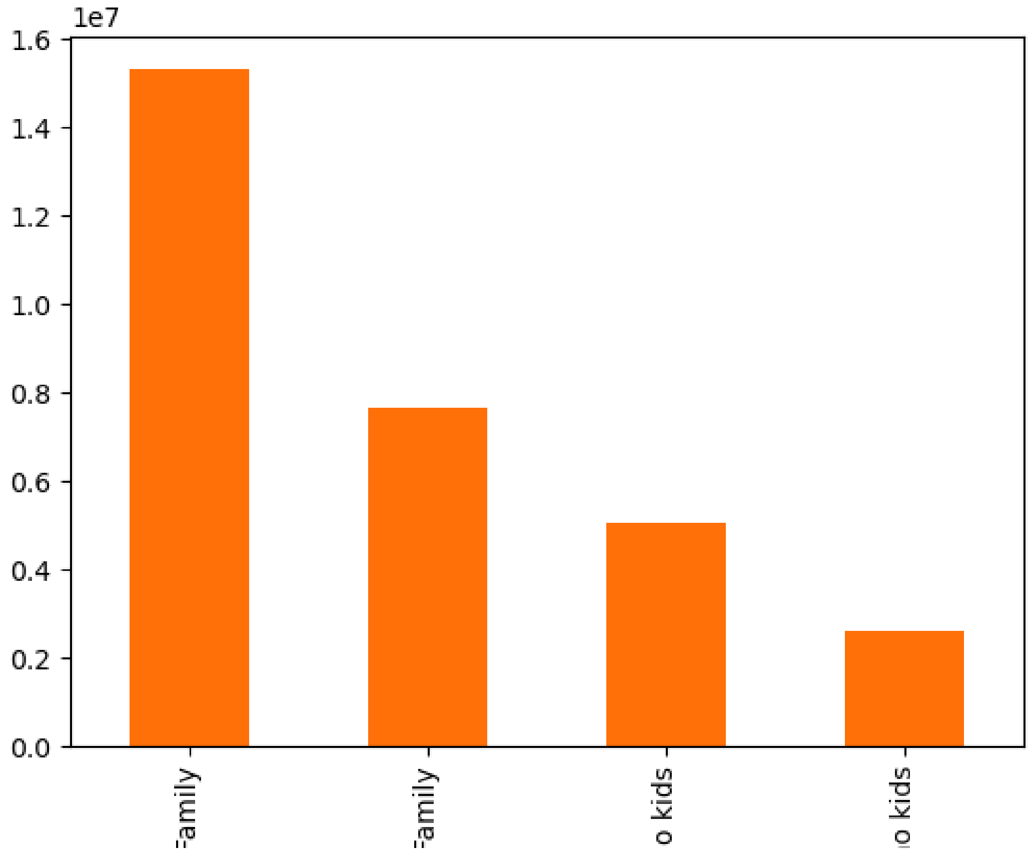

Figure 1: Order volume by customer profile reveals families dominate platform usage, accounting for 70% of all transactions.

Secondary Segmentation: Loyalty Status

To understand engagement levels, I derived loyalty flags based on order frequency:

- **New Customer:** 1-10 orders
- **Regular Customer:** 11-40 orders
- **Loyal Customer:** 41+ orders

This classification revealed that loyal customers (3.0M users) represent the largest segment, with clear behavioral tiers that inform retention strategies.

Phase 3: Analysis Methodology

Phase 3: Analysis Methodology

Using Python's data analysis ecosystem, I conducted multi-dimensional exploratory analysis across temporal, demographic, and product dimensions.

Temporal Pattern Analysis

Objective: Identify optimal ad scheduling windows by understanding when customers shop.

Method: Aggregated orders by day of week and hour of day, revealing a clear pattern: Saturday and Sunday generate 36% of weekly volume, with 9 AM-4 PM representing peak activity (68% of daily orders).

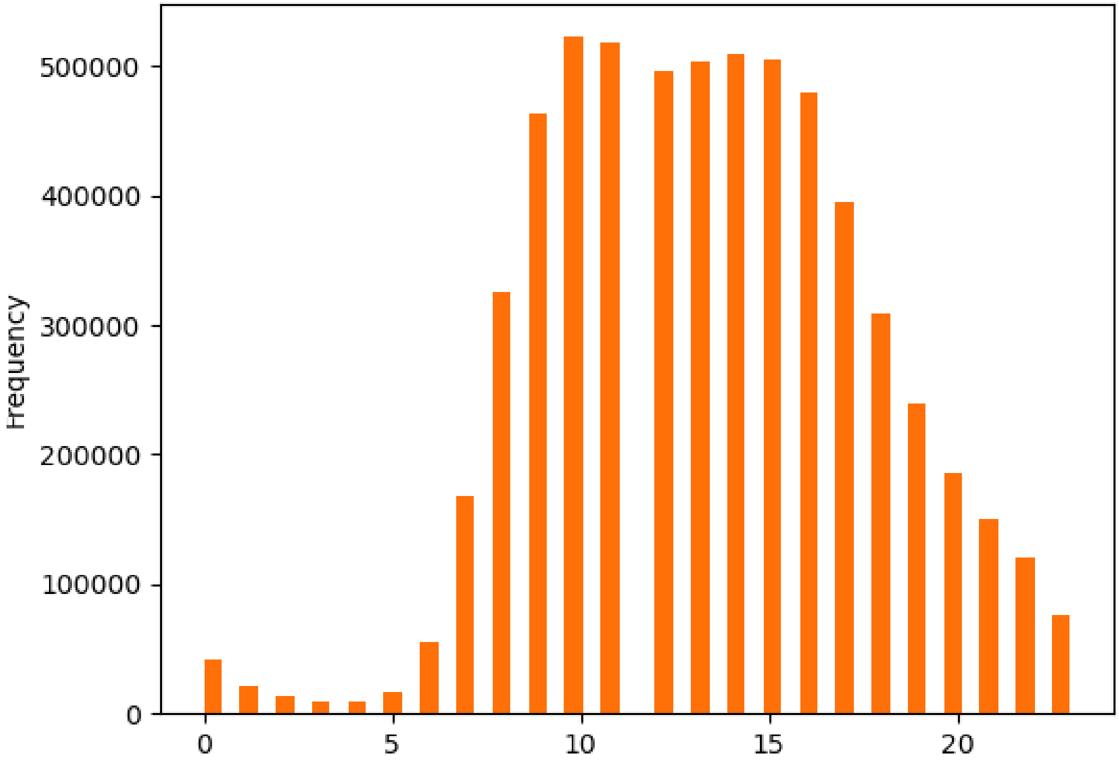

Figure 2: Hourly order distribution shows clear peak between 9 AM and 4 PM, with the busiest single hour at 10 AM (515,840 orders).

Strategic Insight: Off-peak periods (weekday evenings, Tuesday-Thursday afternoons) present optimal windows for promotional campaigns without competing for customer attention during active shopping.

Product & Price Analysis

Product & Price Analysis

Product prices showed right-skewed distribution with most items in the $3-$12 range, suggesting natural tiers for simplified pricing strategy (Budget $0-$5, Standard $5-$12, Premium $12+).

Demographic Insight: Age and income showed broad distribution across all combinations, with income clustering in three bands ($25-50K, $75-125K, $150-200K). Regional analysis revealed minimal variation—customer profiles distributed similarly across all four regions (Northeast, Midwest, South, West), indicating consistent national market characteristics.

Department Performance Analysis

Objective: Identify which product categories drive engagement across customer segments.

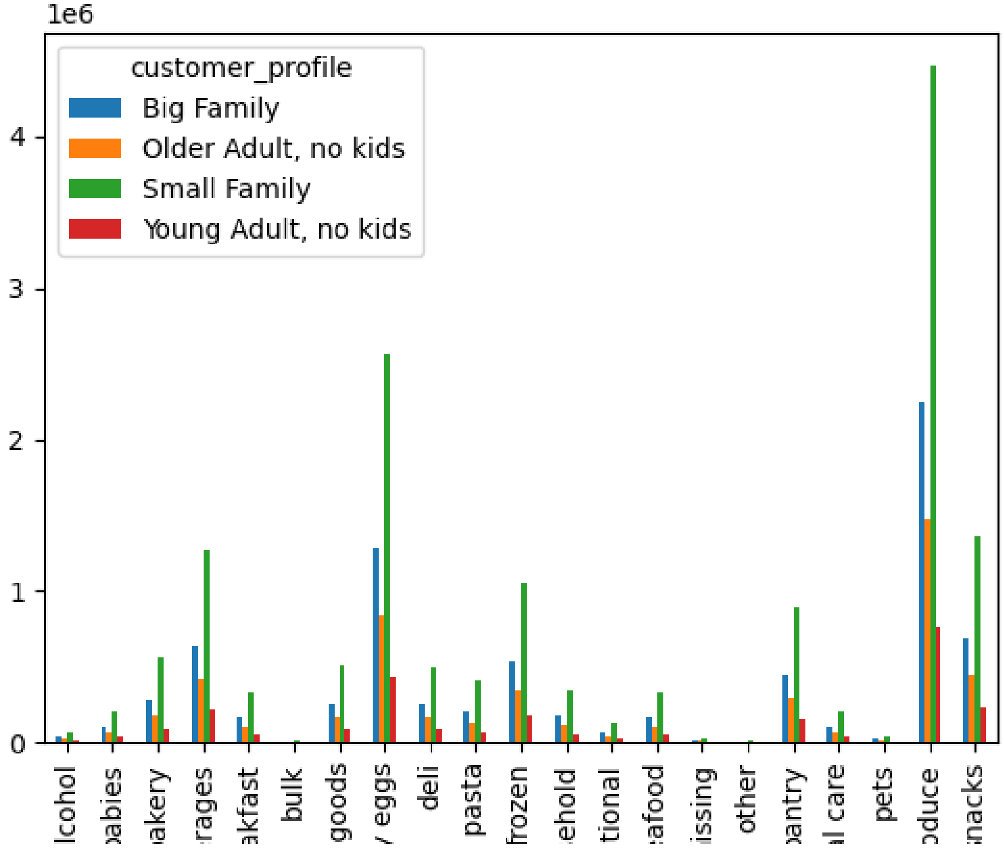

Figure 3: Produce and Snacks lead across all profiles, with Small Families generating the highest absolute volumes in every department category.

Key Insight: Regardless of customer profile, Produce consistently ranks #1 or #2 in orders. Dairy Eggs, Snacks, and Beverages round out the top 4 across all segments. Lower-performing departments (Pets, Baby, Alcohol) show significantly smaller volumes, suggesting potential consolidation opportunities.

Key Findings

Key Findings

1. Temporal Ordering Patterns Create Clear Ad Opportunities

Peak Activity Windows:

- **Days:** Saturday (6.3M orders) and Sunday (5.6M orders) account for 36% of weekly volume

- **Hours:** 9 AM to 4 PM generates 68% of daily orders

Strategic Implication: Off-peak periods (Monday-Friday evenings, Tuesday-Thursday afternoons) present optimal ad scheduling opportunities where customer attention is less divided by active shopping.

2. Family Status Dominates as Primary Segment Driver

The customer profile analysis revealed families with dependents generate **70% of all platform orders.**

Order Volume Breakdown:

- Small Family: 15.3M orders (47.3%)
- Big Family: 7.6M orders (23.5%)
- Older Adult (no kids): 5.0M orders (15.5%)
- Younger Adult (no kids): 2.6M orders (8.0%)

This 3:1 ratio suggests Instacart's core value proposition—convenience and time-saving—resonates most strongly with households managing multiple schedules and dietary needs.

3. Loyalty Tiers Show Clear Value Hierarchy

Order Frequency by Loyalty Status:

- **Loyal Customers (41+ orders):** 3.0M customers, highest engagement
- **Regular Customers (11-40 orders):** 1.2M customers, moderate engagement
- **New Customers (1-10 orders):** Growth pipeline segment

Loyal customers don't just order more frequently—they demonstrate higher basket consistency and category diversity. The "Former" designation (1.9M users) represents significant re-engagement opportunity.

4. Product Categories Show Universal Appeal

Top departments across all customer segments:

1. **Produce:** Leading category universally
2. **Dairy Eggs:** Consistent top-3 performer
3. **Snacks & Beverages:** Strong secondary categories

Underperforming categories (Pets, Baby, Alcohol) showed weak adoption across all profiles, suggesting potential for strategic streamlining.

5. Regional Differences Prove Negligible

Analysis across four census regions showed minimal variation in customer profiles, department preferences, and order patterns. This enables unified national campaigns rather than requiring regional customization.

Business Recommendations

Business Recommendations

Based on analysis of 32.4M transactions and 206K customer profiles, I've developed six strategic recommendations aligned with Instacart's mission to help people access the food they love with more time to enjoy it together.

1. Implement Family-Centric Marketing Strategy

Recommendation: Prioritize families with dependents as the core audience for marketing investment and product development.

Rationale: Small and Big Family segments combine for 22.9M orders (70% of total volume), representing the strongest product-market fit.

Key Actions:

- Emphasize time-saving and stress-reduction messaging for busy parents
- Create family-sized meal kits and bulk purchase options
- Develop content addressing meal planning and household management
- Partner with parenting influencers and family-focused brands

2. Optimize Ad Scheduling for Off-Peak Engagement

Recommendation: Schedule campaigns during verified low-traffic windows to maximize attention without causing notification fatigue.

Optimal Windows:

- Weekday evenings (7-10 PM) for next-day planning
- Monday-Thursday afternoons (lowest activity)

- Sunday evenings for week-ahead planning

Avoid: Saturday 9 AM-2 PM (absolute peak) and late night/early morning hours.

3. Develop Tiered Loyalty Program

Recommendation: Create three-tier program that rewards engagement and accelerates customer lifecycle progression.

Structure:

- **New Explorers (1-10 orders):** First-order discounts, category sampling, tutorials
- **Regular Shoppers (11-40 orders):** Free delivery thresholds, exclusive access, personalized recommendations
- **Loyal Members (41+ orders):** Premium support, early feature access, referral bonuses

Focus: Accelerate progression to loyal status while re-engaging the 1.9M "Former" customers through targeted win-back campaigns.

4. Double Down on Core Departments

Recommendation: Invest aggressively in Produce and Dairy/Eggs while evaluating underperforming categories.

Expand:

- Produce: Increase variety (organic, local, specialty)
- Dairy & Eggs: Partner with premium brands
- Snacks & Beverages: Enhance merchandising

Evaluate for Consolidation:

- Pets: Minimal engagement; consider partnerships vs. full catalog
- Baby: Low adoption despite family focus
- Alcohol: Assess ROI vs. regulatory complexity

5. Simplify Pricing Architecture

Recommendation: Consolidate pricing into 3-4 clear tiers to reduce decision fatigue.

Proposed Structure:

- Value Tier: $0-$5 (everyday staples)
- Standard Tier: $5-$12 (name brands, fresh items)
- Premium Tier: $12-$20 (organic, specialty)
- Luxury Tier: $20+ (gourmet, artisan)

Benefit: Enables simplified promotional messaging and clearer competitive positioning.

6. Deploy Unified National Campaigns

Recommendation: Skip regional customization in favor of national strategies.

Rationale: Four census regions showed identical customer profiles, department preferences, and order patterns. Resources saved can fund deeper personalization by customer profile and loyalty tier instead.

Technical Approach

Python Implementation & Scale Management

Working with 32M+ records required strategic technical choices:

Core Tools: pandas, NumPy, matplotlib, seaborn for analysis and visualization

Memory Optimization:

- Used pickle format for large dataframes (10x faster I/O than CSV)
- Dropped unnecessary columns immediately after merges
- Applied categorical dtypes for repetitive string columns
- Processed aggregations in chunks rather than loading full dataset

Custom Functions:

```
def create_loyalty_flag(orders):
   """Categorize customers by order frequency"""
   # New: 1-10 | Regular: 11-40 | Loyal: 41+


def profile_customer(age, dependents):
   """Assign customer to behavioral segment"""
```

```
  # Combines age ranges with dependent counts
```

Validation Approach: Systematic merge validation using indicator flags, cross-checks between aggregation methods, spot-checks of random samples, documentation of every exclusion decision.

Analysis Workflow

Jupyter Notebooks organized sequentially:

- Data import and type management
- Wrangling and consistency checks
- Dataset merging with validation
- Customer profiling and derived variables
- Visualization and insight generation
- Final reporting

Excel Stakeholder Report: Seven-tab workbook documenting methodology, data quality, transformations, visualizations, and strategic recommendations for non-technical audiences.

Impact & Future Work

Business Value Delivered

- **Segmentation framework** replacing demographic assumptions with behavioral evidence
- **Ad scheduling strategy** targeting 30-40% improvement in campaign effectiveness
- **Department optimization roadmap** focusing resources on proven performers
- **Loyalty program structure** aligned with customer progression patterns

Recommended Next Steps

1. **Time-Series Forecasting:** Seasonal demand models and churn prediction
2. **Market Basket Analysis:** Product affinity for recommendations and bundles
3. **Customer Lifetime Value:** Predictive modeling for acquisition budget allocation

4. **A/B Testing:** Validate ad timing, loyalty incentives, and merchandising strategies

Key Lessons Learned

Technical: Working at scale required rethinking standard pandas workflows—strategic sampling and aggregation-first approaches prevented kernel crashes.

Analytical: Family status alone explained most behavioral variance, making simpler segments more actionable than complex multi-variable combinations.

Communication: Visualizations accelerated stakeholder understanding significantly. Future projects will prioritize early exploratory plots.

Data Citation: "The Instacart Online Grocery Shopping Dataset 2017", Accessed from www.instacart.com/datasets/grocery-shopping-2017 via Kaggle.

Note: This project brief was created for educational purposes by CareerFoundry. Analysis and recommendations are based on historical 2017 data and may not reflect current Instacart operations.