# Phase 1: Problem Definition, Research Question, Hypothesis Formulation, and Data Collection

## Problem Definition

### Business Context

Instacart is an online grocery delivery service that operates through an app, connecting customers with personal shoppers who select and deliver groceries from local stores. As the company grows, understanding customer purchasing patterns becomes increasingly critical for business success.

### Problem Statement

Instacart faces challenges in optimizing product recommendations, inventory management, and marketing strategies due to incomplete understanding of customer purchasing patterns. Specifically, the company needs to identify which products are likely to be reordered by customers and understand the relationship between product categories, ordering time patterns, and customer reordering behavior.

### Business Impact

Solving this problem would allow Instacart to:

- Improve product recommendation algorithms to enhance customer experience
- Optimize inventory management based on predicted demand
- Design targeted marketing campaigns for specific customer segments
- Increase basket size by suggesting complementary products
- Reduce customer churn by better anticipating customer needs

## Research Questions

### Primary Research Question

**How can we predict which products a customer will reorder in their next purchase based on their historical ordering patterns, product characteristics, and temporal factors?**

### Secondary Research Questions

1. What product combinations are most frequently purchased together, and how can this information enhance cross-selling strategies?
2. How do temporal factors (day of week, hour of day) influence purchasing patterns across different product categories?
3. What is the relationship between the time elapsed since the previous order and the likelihood of product reordering?
4. How do product characteristics (department, aisle) relate to reordering frequency?
5. Can customer segments be identified based on their product preferences and reordering behaviors?

## Relevance and Importance of the Research

### Market Relevance

The online grocery market is projected to grow substantially in the coming years. Instacart's ability to understand and predict customer behavior will be a key differentiator in this increasingly competitive space.

### Business Value

This research directly addresses core business objectives:

- **Revenue Growth**: Better recommendations can increase basket size and purchase frequency
- **Cost Reduction**: Improved inventory management reduces waste and stockouts
- **Customer Satisfaction**: Anticipating customer needs creates a more seamless shopping experience
- **Competitive Advantage**: Sophisticated analytical capabilities provide an edge over competitors

### Technical Innovation

The research combines market basket analysis with temporal patterns and product hierarchies, representing a comprehensive approach to understanding online grocery shopping behavior.

## Hypothesis Formulation

### Hypothesis 1: Product Reordering Patterns

**H1**: Customers are significantly more likely to reorder products from specific aisles (e.g., produce, dairy) compared to others (e.g., specialty items, seasonal products).
**H0**: There is no significant difference in reordering probability across different aisles.

### Hypothesis 2: Temporal Purchasing Patterns

**H2**: The day of the week and hour of day significantly influence the types of products customers purchase.
**H0**: There is no significant relationship between ordering time and product categories purchased.

### Hypothesis 3: Order Interval and Basket Composition

**H3**: The time elapsed since a customer's previous order is a significant predictor of their basket size and composition.
**H0**: Order interval has no significant relationship with basket characteristics.

### Hypothesis 4: Product Co-occurrence

**H4**: Certain product pairs co-occur in customers' baskets at a rate significantly higher than would be expected by random chance.
**H0**: Product co-occurrence follows expected patterns based on individual product popularity.

### Hypothesis 5: Customer Segmentation

**H5**: Distinct customer segments exist based on shopping behavior, and these segments exhibit significantly different reordering patterns.
**H0**: Customer reordering behavior is homogeneous across the user base.

## Data Collection

## Database Overview

The Instacart dataset contains anonymized data on customer orders, with the following tables:

- `aisles`: Store aisle information (aisle_id, aisle)
- `departments`: Store department information (department_id, department)
- `products`: Product details (product_id, product_name, aisle_id, department_id)
- `orders`: Order metadata (order_id, user_id, eval_set, order_number, order_dow, order_hour_of_day, days_since_prior_order)
- `order_products__prior`: Products in prior orders (order_id, product_id, add_to_cart_order, reordered)
- `order_products__train`: Products in training orders (same structure as prior)

## SQL Queries for Data Extraction

### 1. Product Reordering Rates by Department and Aisle

```
SELECT
    d.department,
    a.aisle,
    COUNT(op.product_id) AS total_orders,
    SUM(op.reordered) AS reorder_count,
    ROUND(SUM(op.reordered) * 1.0 / COUNT(op.product_id), 4) AS reorder_rate
FROM order_products__prior op
JOIN products p ON op.product_id = p.product_id
JOIN aisles a ON p.aisle_id = a.aisle_id
JOIN departments d ON p.department_id = d.department_id
GROUP BY d.department, a.aisle
ORDER BY reorder_rate DESC;
```

### 2. Order Distribution by Day of Week and Hour

```
SELECT
    order_dow,
    order_hour_of_day,
    COUNT(order_id) AS order_count,
    ROUND(COUNT(order_id) * 100.0 / (SELECT COUNT(*) FROM orders), 2) AS
percentage
FROM orders
GROUP BY order_dow, order_hour_of_day
ORDER BY order_dow, order_hour_of_day;
```

### 3. Impact of Days Since Prior Order on Basket Size

```
SELECT
    CASE
        WHEN days_since_prior_order IS NULL THEN 'First Order'
```

```sql
        WHEN days_since_prior_order < 7 THEN 'Less than a week'
        WHEN days_since_prior_order < 14 THEN '1-2 weeks'
        WHEN days_since_prior_order < 21 THEN '2-3 weeks'
        WHEN days_since_prior_order < 28 THEN '3-4 weeks'
        ELSE 'More than 4 weeks'
    END AS order_interval,
    COUNT(DISTINCT o.order_id) AS num_orders,
    ROUND(AVG(basket_size), 2) AS avg_basket_size,
    ROUND(AVG(reorder_ratio), 4) AS avg_reorder_ratio
FROM orders o
JOIN (
    SELECT
        order_id,
        COUNT(product_id) AS basket_size,
        SUM(reordered) * 1.0 / COUNT(product_id) AS reorder_ratio
    FROM order_products__prior
    GROUP BY order_id
) bs ON o.order_id = bs.order_id
GROUP BY order_interval
ORDER BY
    CASE order_interval
        WHEN 'First Order' THEN 0
        WHEN 'Less than a week' THEN 1
        WHEN '1-2 weeks' THEN 2
        WHEN '2-3 weeks' THEN 3
        WHEN '3-4 weeks' THEN 4
        ELSE 5
    END;
```

#### 4. Frequently Co-purchased Products

```sql
WITH product_pairs AS (
    SELECT
        a.product_id AS product_1,
        b.product_id AS product_2,
        COUNT(*) AS pair_count
    FROM order_products__prior a
    JOIN order_products__prior b ON a.order_id = b.order_id AND a.product_id < b.product_id
    GROUP BY a.product_id, b.product_id
    HAVING COUNT(*) > 100
)
SELECT
    p1.product_name AS product_1_name,
    p2.product_name AS product_2_name,
    pp.pair_count
FROM product_pairs pp
JOIN products p1 ON pp.product_1 = p1.product_id
JOIN products p2 ON pp.product_2 = p2.product_id
ORDER BY pp.pair_count DESC
LIMIT 20;
```

### 5. Customer Segmentation by Shopping Behavior

```sql
WITH user_metrics AS (
    SELECT
        o.user_id,
        COUNT(DISTINCT o.order_id) AS total_orders,
        AVG(basket.basket_size) AS avg_basket_size,
        AVG(basket.reorder_ratio) AS avg_reorder_ratio,
        AVG(o.days_since_prior_order) AS avg_days_between_orders,
        SUM(CASE WHEN o.order_dow IN (0, 6) THEN 1 ELSE 0 END) * 1.0 /
            COUNT(o.order_id) AS weekend_order_ratio
    FROM orders o
    JOIN (
        SELECT
            order_id,
            COUNT(product_id) AS basket_size,
            SUM(reordered) * 1.0 / COUNT(product_id) AS reorder_ratio
        FROM order_products__prior
        GROUP BY order_id
    ) basket ON o.order_id = basket.order_id
    WHERE o.eval_set = 'prior'
    GROUP BY o.user_id
    HAVING COUNT(DISTINCT o.order_id) >= 5
)
SELECT
    CASE
        WHEN avg_basket_size <= 5 THEN 'Small Basket'
        WHEN avg_basket_size <= 12 THEN 'Medium Basket'
        ELSE 'Large Basket'
    END AS basket_segment,
    CASE
        WHEN avg_days_between_orders <= 7 THEN 'Frequent Shopper'
        WHEN avg_days_between_orders <= 14 THEN 'Regular Shopper'
        ELSE 'Occasional Shopper'
    END AS frequency_segment,
    CASE
        WHEN avg_reorder_ratio <= 0.3 THEN 'Low Reorder'
        WHEN avg_reorder_ratio <= 0.6 THEN 'Medium Reorder'
        ELSE 'High Reorder'
    END AS reorder_segment,
    COUNT(*) AS user_count,
    ROUND(AVG(total_orders), 1) AS avg_orders,
    ROUND(AVG(avg_basket_size), 1) AS avg_items_per_order,
    ROUND(AVG(avg_reorder_ratio), 3) AS avg_reorder_rate,
    ROUND(AVG(avg_days_between_orders), 1) AS avg_days_between_orders,
    ROUND(AVG(weekend_order_ratio), 3) AS weekend_order_rate
FROM user_metrics
GROUP BY basket_segment, frequency_segment, reorder_segment
ORDER BY user_count DESC;
```

# Data Collection Strategy

The data collection strategy leverages the Instacart database to extract relevant information for hypothesis testing:

1. **Comprehensive Database**: The analysis utilizes the complete Instacart dataset, which includes over 3 million orders from more than 200,000 users.
2. **SQL-Based Extraction**: Advanced SQL queries are used to transform raw order data into analytical datasets that support hypothesis testing.
3. **Multi-Table Integration**: Data is collected by joining multiple tables to connect product characteristics with ordering patterns and customer behavior.
4. **Feature Aggregation**: Order-level data is aggregated to create user-level features that capture behavioral patterns over time.
5. **Temporal Analysis**: Special attention is given to temporal features like order day, time, and interval between orders.
6. **Data Segmentation**: The collection strategy facilitates segmentation of products, time periods, and customers to enable nuanced analysis.