# Instacart

## Market Basket Analysis

Which products will an Instacart consumer purchase again?

Boriss Siliverstovs

July 21, 2017

# Roadmap

- Motivation
- Data
- Results
- Conclusions

# Motivation: What do I do?

Running Kaggle competition
(August 14, 2017 - Final submission deadline)

- Data are provided by Instacart
  - Online retail grocery shop in US
  - Founded in 2012
  - Present in 1,200 cities in 25 states (March 2017)
  - Valued at approximately 3.4 bln USD (March 2017)
- Data contain ordering history of Instacart customers
  - *Questions*: Which products will a consumer
    - add to their cart during the next order?
    - try for the first time?
    - purchase again?
- **Given a sequence of prior orders**
  **which products will be re-ordered in the next order?**

# Motivation: What do I do?

Running Kaggle competition
(August 14, 2017 - Final submission deadline)

- Data are provided by Instacart
  - Online retail grocery shop in US
  - Founded in 2012
  - Present in 1,200 cities in 25 states (March 2017)
  - Valued at approximately 3.4 bln USD (March 2017)

- Data contain ordering history of Instacart customers
  - *Questions*: Which products will a consumer
    - add to their cart during the next order?
    - try for the first time?
    - purchase again?

- **Given a sequence of prior orders**
  **which products will be re-ordered in the next order?**

# Motivation: What do I do?

Running Kaggle competition
(August 14, 2017 - Final submission deadline)

- Data are provided by Instacart
  - Online retail grocery shop in US
  - Founded in 2012
  - Present in 1,200 cities in 25 states (March 2017)
  - Valued at approximately 3.4 bln USD (March 2017)

- Data contain ordering history of Instacart customers
  - *Questions*: Which products will a consumer
    - add to their cart during the next order?
    - try for the first time?
    - purchase again?

- **Given a sequence of prior orders**
  **which products will be re-ordered in the next order?**

# Motivation: What do I do?

Running Kaggle competition
(August 14, 2017 - Final submission deadline)

- Data are provided by Instacart
  - Online retail grocery shop in US
  - Founded in 2012
  - Present in 1,200 cities in 25 states (March 2017)
  - Valued at approximately 3.4 bln USD (March 2017)

- Data contain ordering history of Instacart customers
  - *Questions*: Which products will a consumer
    - add to their cart during the next order?
    - try for the first time?
    - purchase again?

- **Given a sequence of prior orders**
  **which products will be re-ordered in the next order?**

# Motivation: What do I do?

Running Kaggle competition
(August 14, 2017 - Final submission deadline)

- Data are provided by Instacart
  - Online retail grocery shop in US
  - Founded in 2012
  - Present in 1,200 cities in 25 states (March 2017)
  - Valued at approximately 3.4 bln USD (March 2017)

- Data contain ordering history of Instacart customers
  - *Questions*: Which products will a consumer
    - add to their cart during the next order?
    - try for the first time?
    - purchase again?
- **Given a sequence of prior orders**
  **which products will be re-ordered in the next order?**

# Motivation: What do I do?

Running Kaggle competition
(August 14, 2017 - Final submission deadline)

- Data are provided by Instacart
  - Online retail grocery shop in US
  - Founded in 2012
  - Present in 1,200 cities in 25 states (March 2017)
  - Valued at approximately 3.4 bln USD (March 2017)

- Data contain ordering history of Instacart customers
  - *Questions*: Which products will a consumer
    - add to their cart during the next order?
    - try for the first time?
    - purchase again?
- **Given a sequence of prior orders**
  **which products will be re-ordered in the next order?**

# Motivation: What do I do?

Running Kaggle competition
(August 14, 2017 - Final submission deadline)

- Data are provided by Instacart
  - Online retail grocery shop in US
  - Founded in 2012
  - Present in 1,200 cities in 25 states (March 2017)
  - Valued at approximately 3.4 bln USD (March 2017)

- Data contain ordering history of Instacart customers
  - *Questions*: Which products will a consumer
    - add to their cart during the next order?
    - try for the first time?
    - purchase again?

- **Given a sequence of prior orders**
  **which products will be re-ordered in the next order?**

# Kaggle challenge: Predict re-ordered products

Table: A typical consumer order

| Product | Reordered |
|---|---|
| Chocolate Sandwich Cookies | 1 |
| Dry Nose Oil | 0 |
| Pure Coconut Water With Orange | 1 |
| Green Chile Anytime Sauce | 0 |

Table: (Correct) prediction

| Product | Reordered |
|---|---|
| Chocolate Sandwich Cookies | 1 |
| Pure Coconut Water With Orange | 1 |

# Motivation: Who cares?

(Online) retailers do

- Complement/substitute shopping lists of customers
  - by sending (accurate) personalised reminders
- Time-saving shopping activity
  - Increase customer turnover
- Better management of inventories
- Allow dynamic pricing
  - by offering personalised price tags/coupons
- Boost sales/revenues/market share

I do:

- The Kaggle competition prize
- Built up a reputation

# Motivation: Who cares?

(Online) retailers do

- Complement/substitute shopping lists of customers
  - by sending (accurate) personalised reminders
- Time-saving shopping activity
  - Increase customer turnover
- Better management of inventories
- Allow dynamic pricing
  - by offering personalised price tags/coupons
- Boost sales/revenues/market share

I do:

- The Kaggle competition prize
- Built up a reputation

# Motivation: Who cares?

(Online) retailers do

- Complement/substitute shopping lists of customers
  - by sending (accurate) personalised reminders
- Time-saving shopping activity
  - Increase customer turnover
- Better management of inventories
- Allow dynamic pricing
  - by offering personalised price tags/coupons
- Boost sales/revenues/market share

I do:

- The Kaggle competition prize
- Built up a reputation

# Motivation: Who cares?

(Online) retailers do

- Complement/substitute shopping lists of customers
  - by sending (accurate) personalised reminders
- Time-saving shopping activity
  - Increase customer turnover
- Better management of inventories
- Allow dynamic pricing
  - by offering personalised price tags/coupons
- Boost sales/revenues/market share

I do:

- The Kaggle competition prize
- Built up a reputation

# Data

- 206,209 customers
- 49,688 products
  - 21 departments
  - 134 aisles
- 3,421,083 orders
  - 3,214,874 orders in the prior set
  - 206,209 orders in the postprior set
    - 131,209 orders in the **train** set
    - 75,000 orders in the **test** set ( stored in Kaggle )

# Data features

- Product-specific
- User-specific
- User-Product-specific
- Train/Test-specific

# Data features: Product-specific

- How many times a product was ordered
- How many times a product was re-ordered
  - Product re-order ratio
- How many times a product was ordered for the first time
- How many times a product was ordered for the second time
  - Re-order propensity of a product

# Data features: User-specific

- Number of orders made by user
- Number of all products ever ordered by user
  - Average basket size
- Number of distinct products ever ordered by user
- DOW/HOD - orders are typically made
- Average ordering frequency in days
- Average share of re-ordered products in each order

# Data features: User-Product-specific

- How many times user ordered product
- Share of orders with a product in total number of orders
- Share of orders with a product since this product was ordered for the first time
- How many orders/days ago a product was ordered last time
- Average position in ordering cart

# Data features: Train/Test-specific

- Days since the last order in prior
- Day of week order was made
- Hour of day order was made

# Kaggle challenge: Predict re-ordered products

Table: A typical consumer order

| Product | Reordered |
|---|---|
| Chocolate Sandwich Cookies | 1 |
| Dry Nose Oil | 0 |
| Pure Coconut Water With Orange | 1 |
| Green Chile Anytime Sauce | 0 |

Table: (Correct) prediction

| Product | Reordered |
|---|---|
| Chocolate Sandwich Cookies | 1 |
| Pure Coconut Water With Orange | 1 |

Ordering habits of a user:

- Basket size
- Share of re-ordered products

Investigate their variation with data features

# Kaggle challenge: Predict re-ordered products

Table: A typical consumer order

| Product | Reordered |
|---|---|
| Chocolate Sandwich Cookies | 1 |
| Dry Nose Oil | 0 |
| Pure Coconut Water With Orange | 1 |
| Green Chile Anytime Sauce | 0 |

Table: (Correct) prediction

| Product | Reordered |
|---|---|
| Chocolate Sandwich Cookies | 1 |
| Pure Coconut Water With Orange | 1 |

Ordering habits of a user:

- Basket size
- Share of re-ordered products

Investigate their variation with data features

# Kaggle challenge: Predict re-ordered products

Table: A typical consumer order

| Product | Reordered |
|---|---|
| Chocolate Sandwich Cookies | 1 |
| Dry Nose Oil | 0 |
| Pure Coconut Water With Orange | 1 |
| Green Chile Anytime Sauce | 0 |

Table: (Correct) prediction

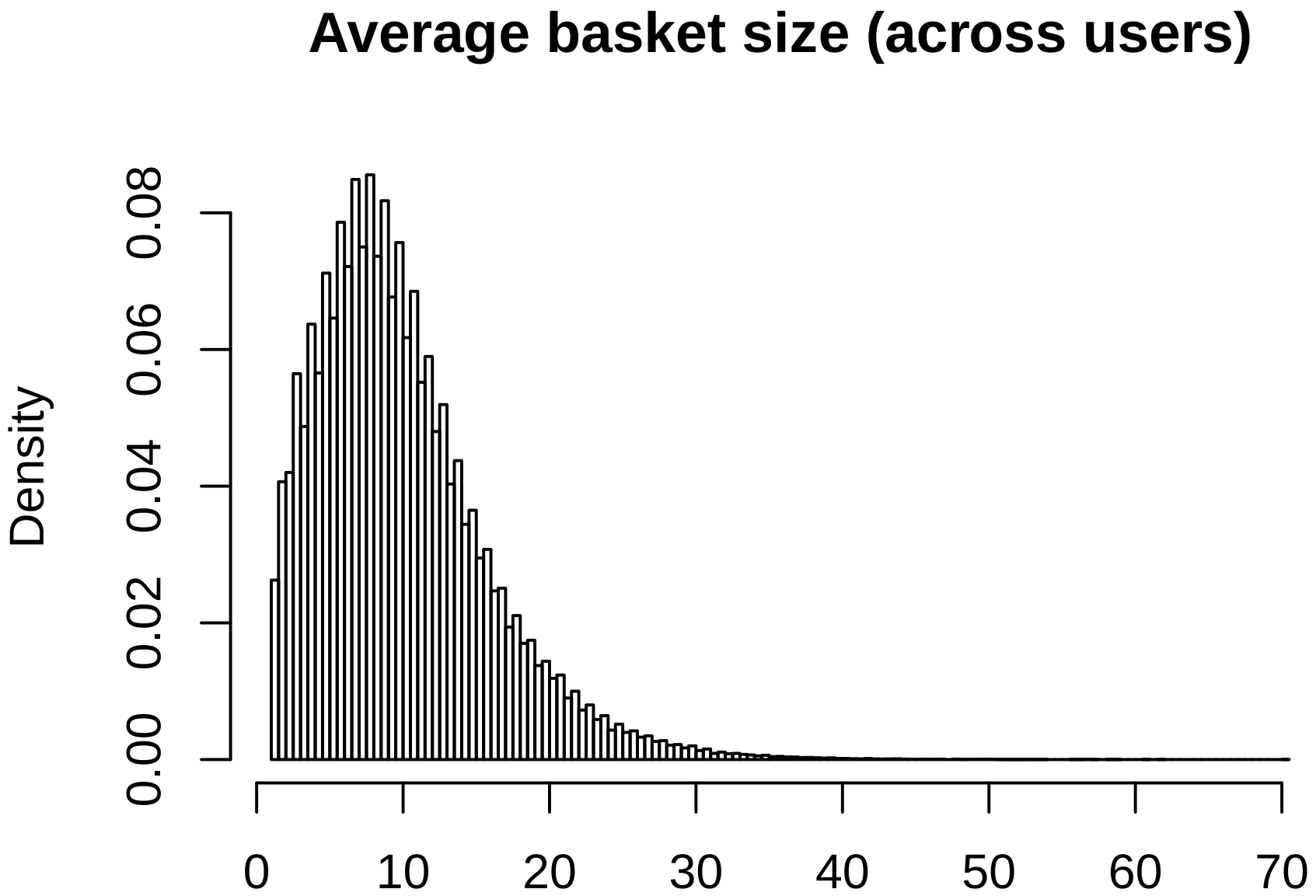
| Product | Reordered |
|---|---|
| Chocolate Sandwich Cookies | 1 |
| Pure Coconut Water With Orange | 1 |

Ordering habits of a user:

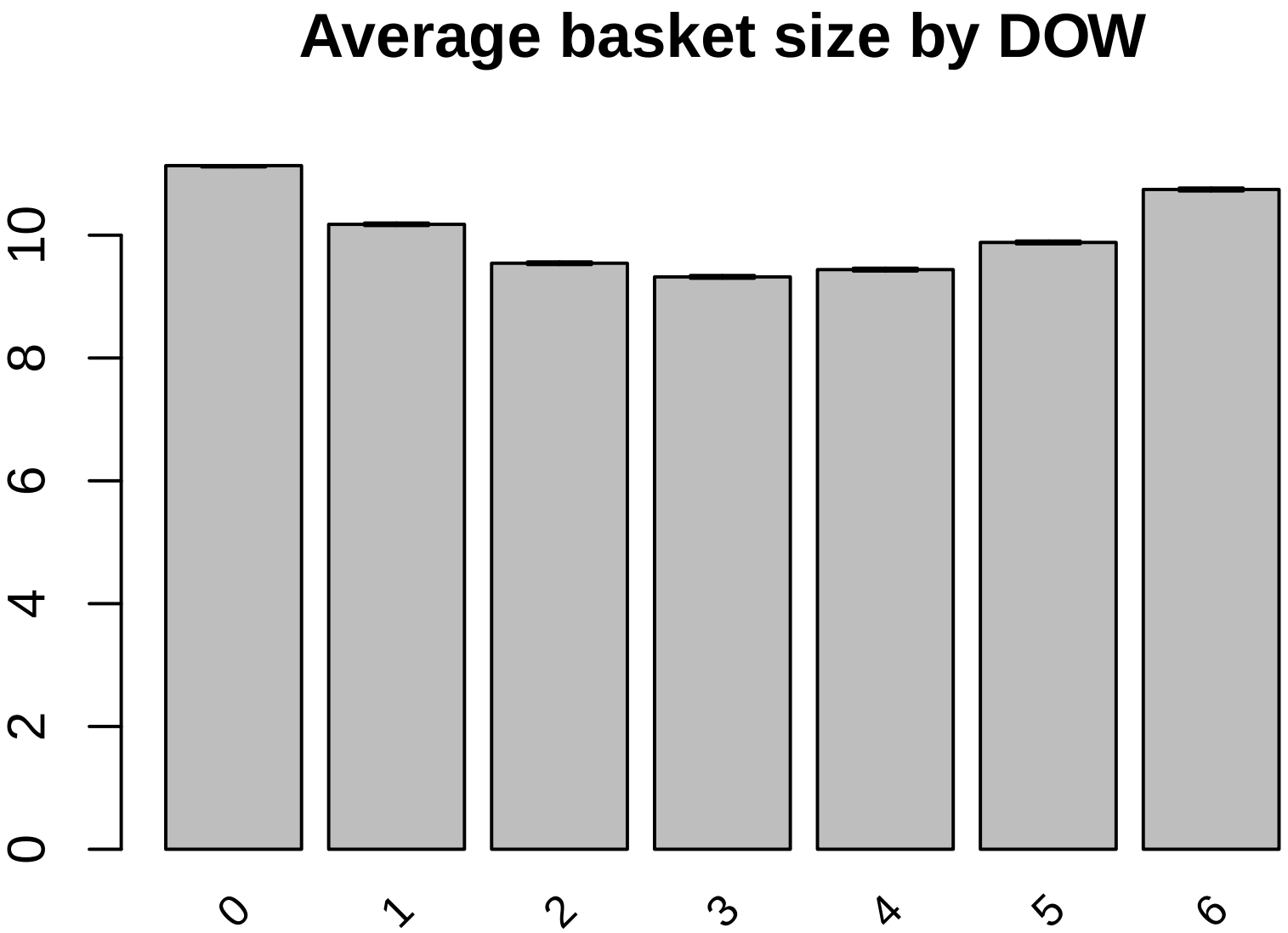
- Basket size
- Share of re-ordered products

Investigate their variation with data features

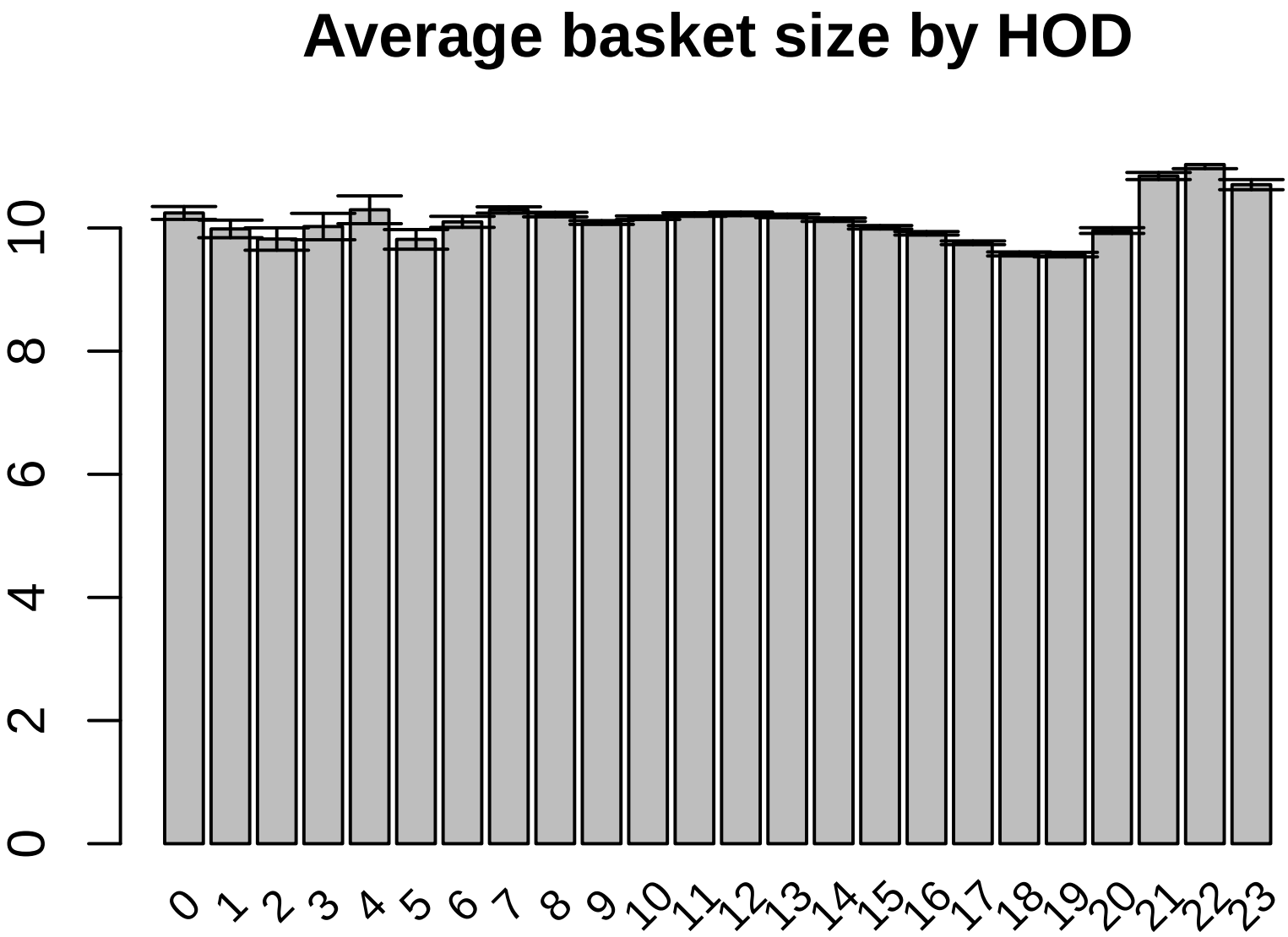
# Data: basket size



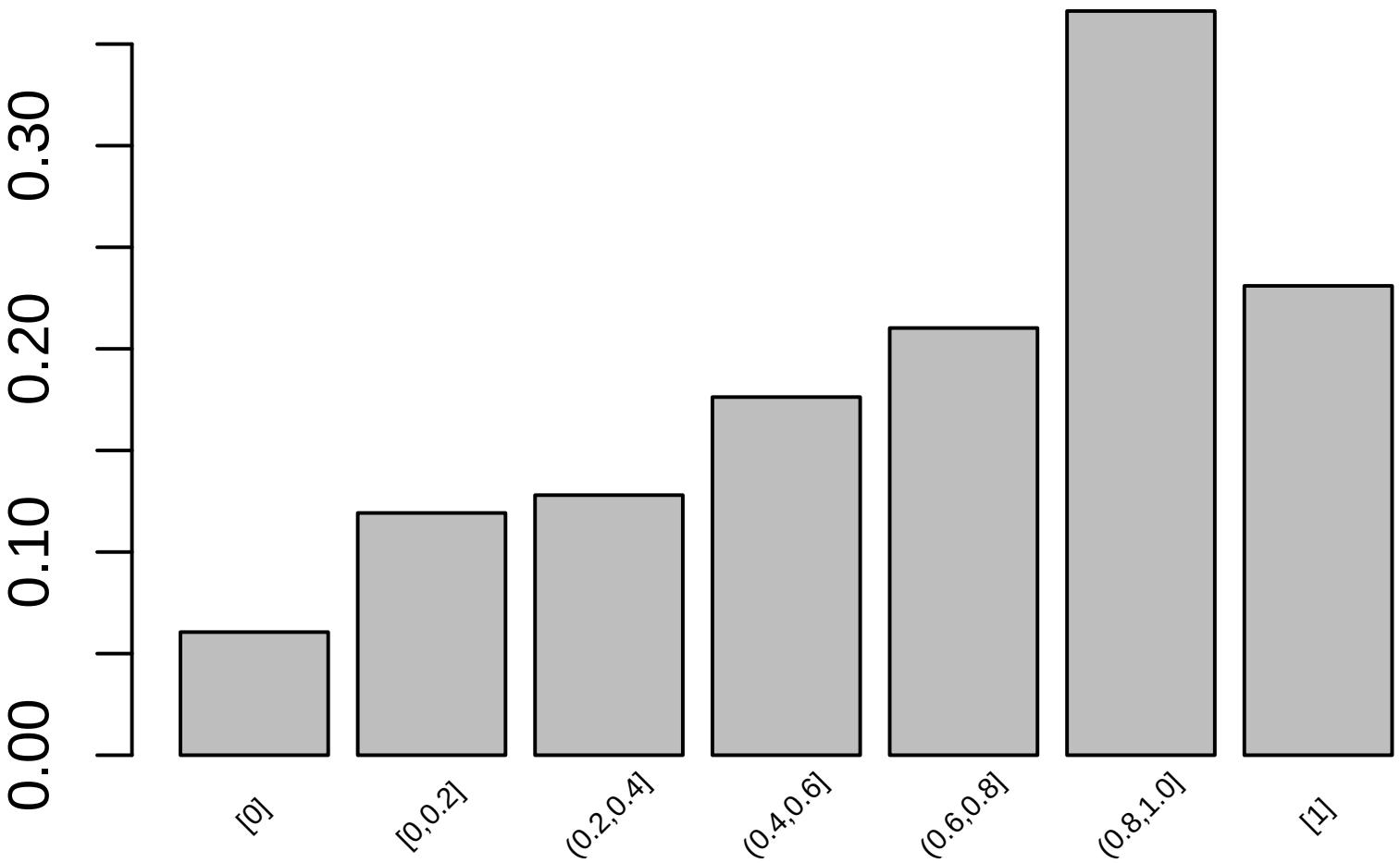
# Data: basket size by DOW



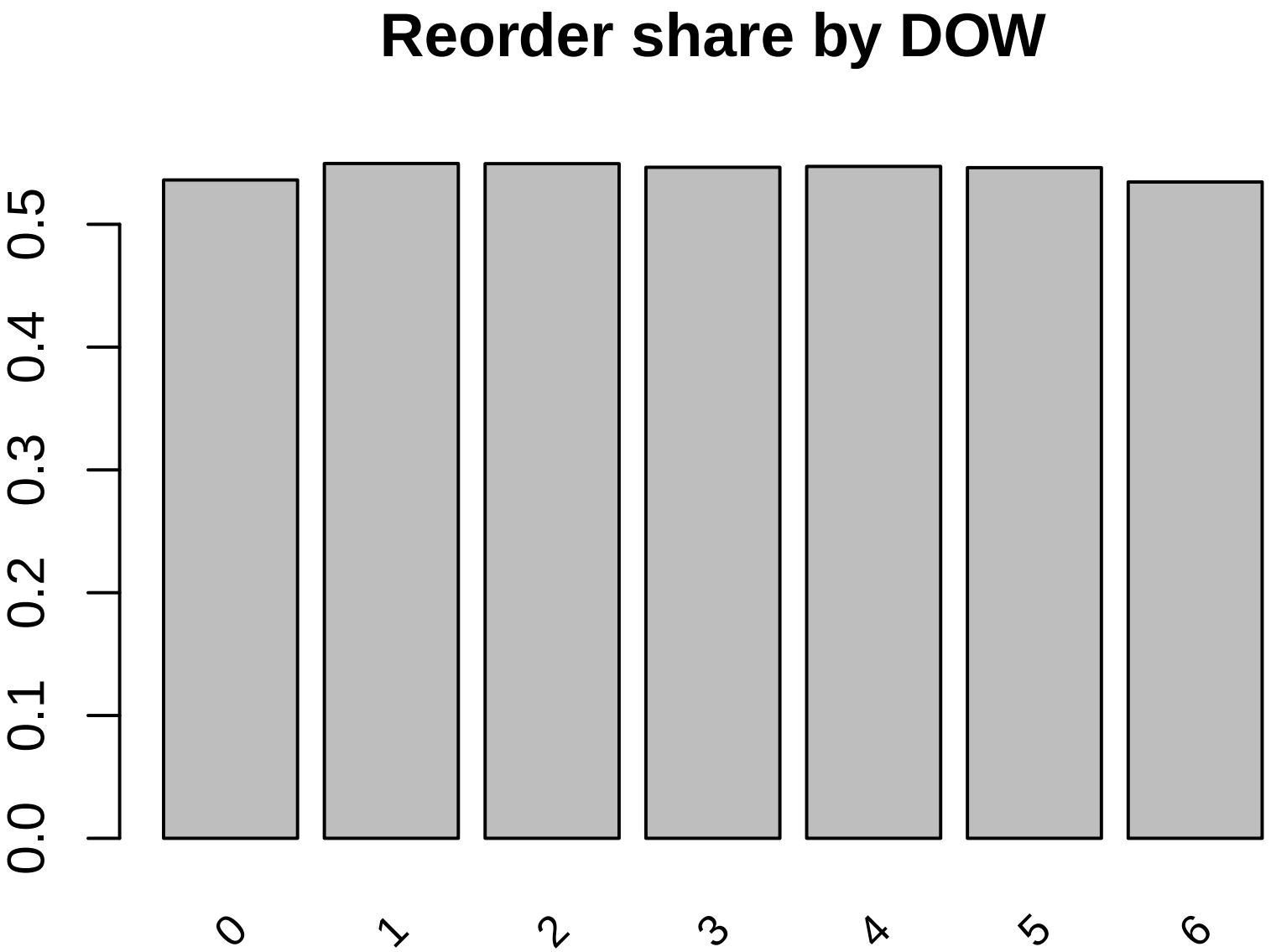
# Data: basket size by HOD



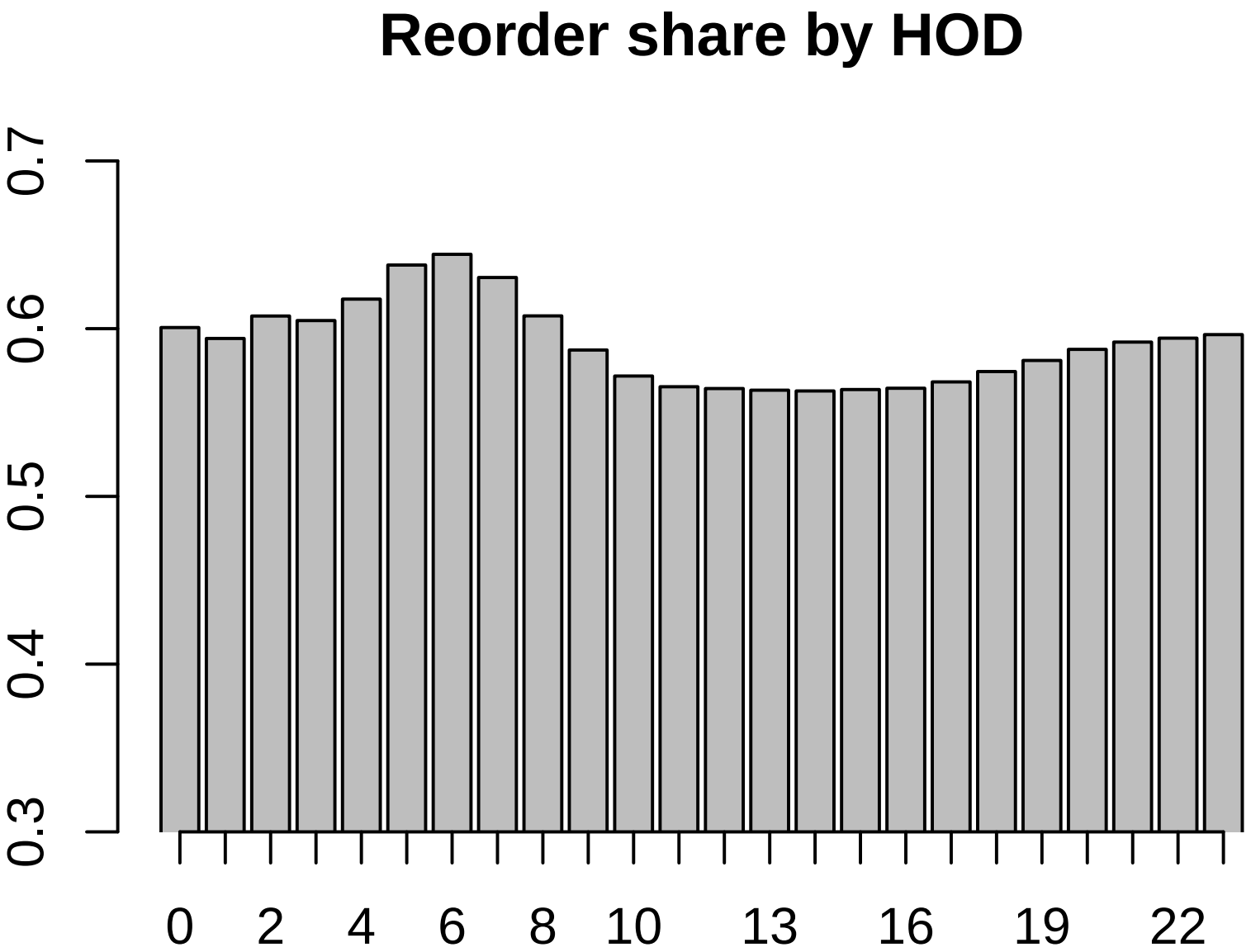
# Data: reorder share



# Data: Reorder share by DOW



# Data: Reorder share by HOD



# Evaluation metrics: mean $F_1$ score

$F_1$ score is the harmonic mean of:

- $p$ - precision: share of correct predictions in total number of predictions, $\widehat{x}$
- $r$ - recall: share of correct predictions in total number of outturns, $x$

$$p(x, \widehat{x}) = \frac{|x \cap \widehat{x}|}{|\widehat{x}|} \qquad r(x, \widehat{x}) = \frac{|x \cap \widehat{x}|}{|x|}$$

$$F_1 score = 2 * \frac{1}{\frac{1}{p} + \frac{1}{r}} = 2 * \frac{p * r}{p + r} \in [0, 1]$$

- At Kaggle: take average of individual $F_1$ scores

# Benchmark models

- Repeat last order
- Repeat reordered items in last order
- Repeat reordered items in two last orders
- **Repeat reordered items in three last orders**
  - Kaggle Leadership Board:

    $F_1$ score $= 0.3586635$ (1,074th place out of 1,716)

    $F_1^{MAX}$ score $= 0.4069657$

# Benchmark models

- Repeat last order
- Repeat reordered items in last order
- Repeat reordered items in two last orders
- **Repeat reordered items in three last orders**
  - Kaggle Leadership Board:

    $F_1$ score $= 0.3586635$ (1,074th place out of 1,716)

    $F_1^{MAX}$ score $= 0.4069657$

# Benchmark models

- Repeat last order (LB = 0.3276746 )
- Repeat reordered items in last order (LB = 0.3276826)
- Repeat reordered items in two last orders
- **Repeat reordered items in three last orders**
  - Kaggle Leadership Board:

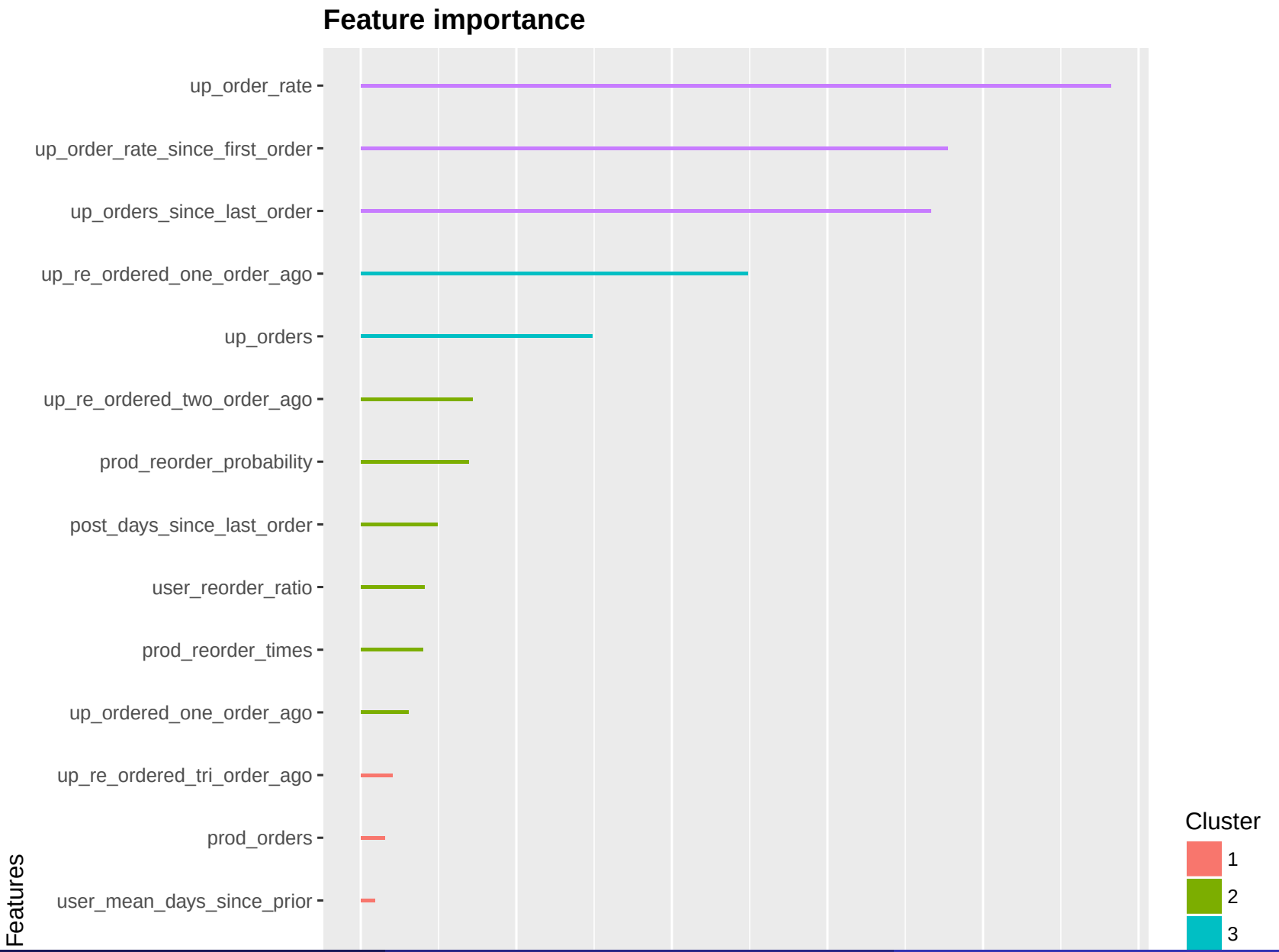
    $F_1$ score = 0.3586635 (1,074th place out of 1,716)

    $F_1^{MAX}$ score = 0.4069657

# XGBoost

- King of Kaggle competitions
  - Praised to be one of the most successful algorithms
- Provides feature importance
  - Product-specific
  - User-specific
  - User-Product-specific
  - Train/Test-specific

# XGBoost


Feature importance
up_order_rate
up_order_rate_since_first_order
up_orders_since_last_order
up_re_ordered_one_order_ago
up_orders
up_re_ordered_two_order_ago
prod_reorder_probability
post_days_since_last_order
user_reorder_ratio
prod_reorder_times
up_ordered_one_order_ago
up_re_ordered_tri_order_ago
prod_orders
user_mean_days_since_prior
Features
Cluster
1
2
3

# XGBoost

- Current stand:
  - My best LB score: 0.3797138 in TOP 33% ( 561 out of 1,716 )
- Outlook
  - Think of more features to add
  - Compare benchmark and XGBoost
    - Predict those cases where benchmark beats XGBoost
  - User-specific thresholds
  - Apply alternative methods and their combinations:
    - Logistic regression
    - Random Forest
    - Gradient Boosting Machine (GBM)
- Not too late to join (Ends on Aug 14)!

# XGBoost

- Current stand:
  - My best LB score: 0.3797138 in TOP 33% ( 561 out of 1,716 )
- Outlook
  - Think of more features to add
  - Compare benchmark and XGBoost
    - Predict those cases where benchmark beats XGBoost
  - User-specific thresholds
  - Apply alternative methods and their combinations:
    - Logistic regression
    - Random Forest
    - Gradient Boosting Machine (GBM)
- Not too late to join (Ends on Aug 14)!

# XGBoost

- Current stand:
  - My best LB score: 0.3797138 in TOP 33% ( 561 out of 1,716 )
- Outlook
  - Think of more features to add
  - Compare benchmark and XGBoost
    - Predict those cases where benchmark beats XGBoost
  - User-specific thresholds
  - Apply alternative methods and their combinations:
    - Logistic regression
    - Random Forest
    - Gradient Boosting Machine (GBM)
- Not too late to join (Ends on Aug 14)!