

Instacart Market Basket Analysis

Whether you shop from meticulously planned grocery lists or let whimsy guide your grazing, our unique food rituals define who we are. Instacart, a grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them. After selecting products through the Instacart app, personal shoppers review your order and do the in-store shopping and delivery for you.

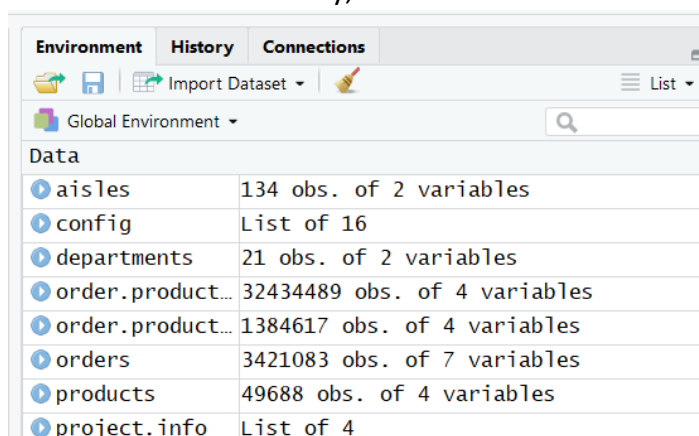
Instacart's data science team plays a big part in providing this delightful shopping experience. Currently they use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session. Recently, Instacart open sourced this data - see their blog post on [3 Million Instacart Orders, Open Sourced](https://www.kaggle.com/c/instacart-market-basket-analysis/overview).

Please refer to this URL for details about this competition:

<https://www.kaggle.com/c/instacart-market-basket-analysis/overview>

Student action:

1. In Rstudio, create a project using ProjectTemplate, details here: http://projecttemplate.net/getting_started.html
2. Unzip the file data.zip and put those files under data directory under the project you just created.
3. Open global.dcf file under config directory, add below libraries after libraries tag and save the file:
reshape2, tidyverse, stringr, lubridate, dplyr, pROC, xgboost, precrec
4. Load the project using load.project() command as stated in ProjectTemplate documentation. You should see all the libraries will be installed and a few dataframes been loaded into memory, like below:



The screenshot shows the RStudio Environment pane with the 'Global Environment' selected. It lists several data frames that have been loaded into memory. The data frames and their dimensions are as follows:

| Variable | Dimensions |
|------------------|------------------------------|
| aisles | 134 obs. of 2 variables |
| config | List of 16 |
| departments | 21 obs. of 2 variables |
| order.product... | 32434489 obs. of 4 variables |
| order.product... | 1384617 obs. of 4 variables |
| orders | 3421083 obs. of 7 variables |
| products | 49688 obs. of 4 variables |
| project.info | List of 4 |

5. Click on each dataframe and make sure you understand their meanings.
6. Join orders dataframe and order.product.prior dataframe by order_id and user_id, please google dplyr on how to join two dataframes.
7. Perform statistical analysis using ggplot and summary function on those dataframes.