Evaluation

```
source("readDataToMemory.R")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## * Using Spark: 2.1.0
```

```
readInstacart()

library(DBI)
library(ggplot2)
library(ggthemes)

src_tbls(sc)
```

```
## [1] "departments_tbl"           "order_products__prior_tbl"
## [3] "order_products__train_tbl" "orders_tbl"
## [5] "products_tbl"
```

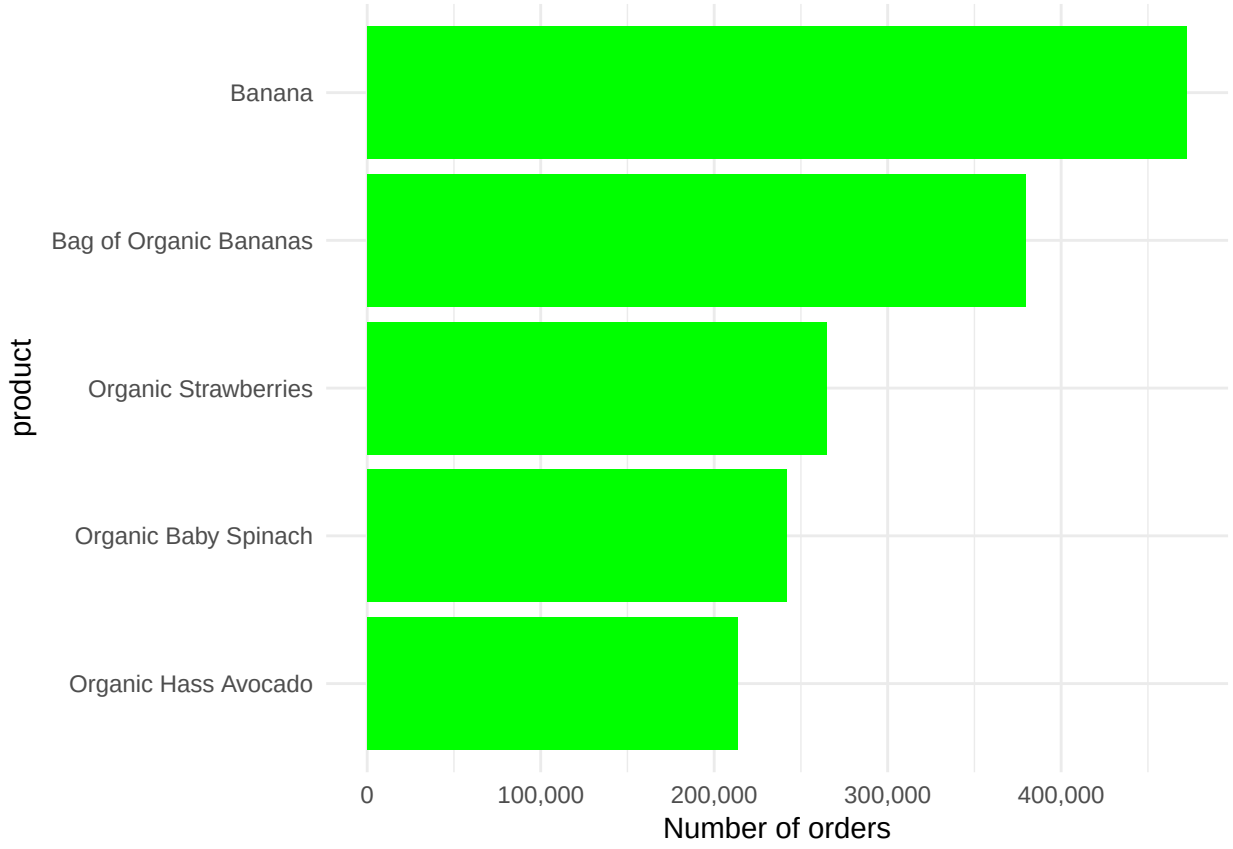
1) The most populars products

```
Mostpopul <- order_products__prior %>%
             group_by(product_id) %>%
             summarize(count = n()) %>%
             top_n(5, wt = count) %>%
             left_join(select(products, product_id, product_name), by="product_id") %>%
             arrange(desc(count)) %>%
             collect()
Mostpopul
```

```
## # A tibble: 5 x 3
##   product_id  count           product_name
##        <int>  <dbl>                  <chr>
## 1      24852 472565                 Banana
## 2      13176 379450 Bag of Organic Bananas
## 3      21137 264683   Organic Strawberries
## 4      21903 241921   Organic Baby Spinach
## 5      47209 213584   Organic Hass Avocado
```

```
Mostpopul %>% ggplot(
    aes(reorder(product_name, count, function(x) x),
        count)) +
  geom_bar(stat="identity", fill='green') +
  coord_flip() +
  scale_y_continuous(label=scales::comma) +
  xlab("product") +
```

```
  ylab("Number of orders") +
  theme_minimal()
```



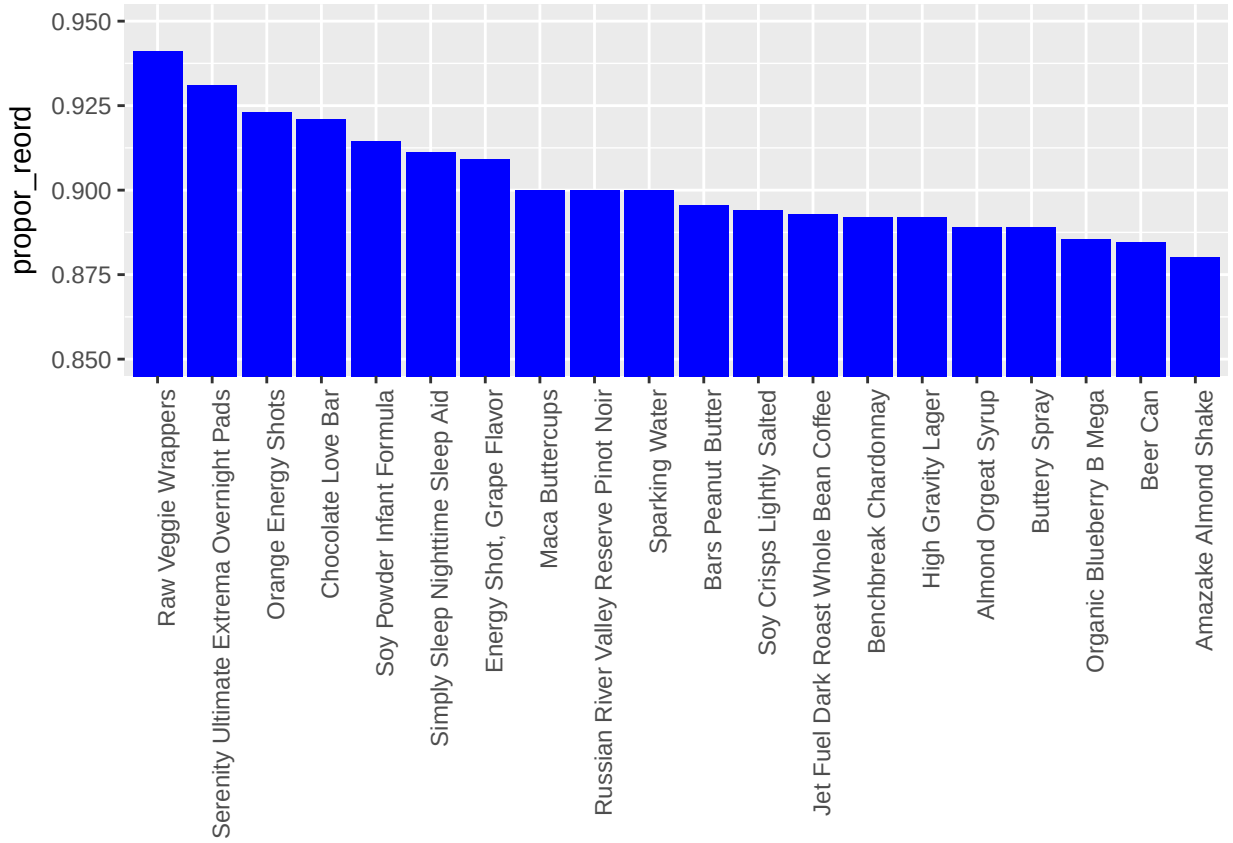
2) The most repeted products

```
Reor <-order_products__prior %>%
       group_by(product_id) %>%
       summarize(propor_reord = mean(reordered), n=n()) %>%
       top_n(20, wt = propor_reord) %>%
       arrange(desc(propor_reord)) %>%
       left_join(select(products, product_id, product_name), by="product_id") %>%
       collect()
Reor
```

```
## # A tibble: 20 x 4
##    product_id propor_reord     n                                product_name
##         <int>        <dbl> <dbl>                                       <chr>
##  1       6433    0.9411765    68                         Raw Veggie Wrappers
##  2       2075    0.9310345    87 Serenity Ultimate Extrema Overnight Pads
##  3      43553    0.9230769    13                         Orange Energy Shots
##  4      27740    0.9207921   101                          Chocolate Love Bar
##  5      14609    0.9142857    35                  Soy Powder Infant Formula
##  6      13875    0.9111111    45          Simply Sleep Nighttime Sleep Aid
##  7      39992    0.9090909    22                  Energy Shot, Grape Flavor
##  8      35604    0.9000000   100                              Maca Buttercups
##  9       5868    0.9000000    30  Russian River Valley Reserve Pinot Noir
## 10      31418    0.9000000    60                              Sparking Water
```

```
## 11      36543    0.8955224    67                      Bars Peanut Butter
## 12      26093    0.8939394    66                Soy Crisps Lightly Salted
## 13        700    0.8928571    28    Jet Fuel Dark Roast Whole Bean Coffee
## 14      38251    0.8918919   111                  Benchbreak Chardonnay
## 15       4212    0.8918919    37                     High Gravity Lager
## 16      38438    0.8888889    27                   Almond Orgeat Syrup
## 17      35513    0.8888889    18                         Buttery Spray
## 18      36801    0.8854167    96                Organic Blueberry B Mega
## 19      34246    0.8846154    52                               Beer Can
## 20      47825    0.8800000    25                   Amazake Almond Shake
```

```
Reor %>%
  ggplot(aes(x=reorder(product_name,-propor_reord), y=propor_reord))+
  geom_bar(stat="identity",fill="blue")+
  theme(axis.text.x=element_text(angle=90, hjust=1),axis.title.x = element_blank())+coord_cartesian(yli
```



3)Product bough in first place

```
Fir <- order_products__prior %>%
       group_by(product_id) %>%
       mutate(add_cart_order = if_else(add_to_cart_order == 1, 1, 0)) %>%
       summarize(first = sum(add_cart_order), n=n()) %>%
       top_n(10, wt = first) %>%
       arrange(desc(first)) %>%
       left_join(select(products, product_id, product_name), by="product_id") %>%
       collect()
Fir
```

```
## # A tibble: 10 x 4
```

```
##    product_id  first      n           product_name
##         <int>  <dbl>  <dbl>                  <chr>
##  1      24852 110916 472565                 Banana
##  2      13176  78988 379450 Bag of Organic Bananas
##  3      27845  30927 137905     Organic Whole Milk
##  4      21137  27975 264683   Organic Strawberries
##  5      47209  24116 213584   Organic Hass Avocado
##  6      21903  23543 241921   Organic Baby Spinach
##  7      47766  22398 176815        Organic Avocado
##  8      19660  16822  56087           Spring Water
##  9      16797  16366 142951           Strawberries
## 10      27966  14393 137057    Organic Raspberries
```

Recommender

```
order_products__prior %>%
  select(order_id, product_id) %>%
  left_join(orders, by="order_id") %>%
  filter(user_id <= 10) %>%
  select(product_id, user_id) %>%
  group_by(user_id, product_id) %>%
  summarise(rating = n()) %>%
  rename(user = user_id) %>%
  mutate(item=product_id) %>%
  select(user, item, rating) ->
  user_item_rating
```

```
explicit_model <- ml_als_factorization( user_item_rating, iter.max = 5, regularization.parameter = 0.01)
```

```
v <- as.matrix(explicit_model$item.factors)[, -1]

u <- as.matrix(explicit_model$user.factors)[, -1]

a <- u %*% t(v)
```

Most recomended product for each person (first 20)

```
Max <- apply(a, 1, which.max)
dat <- as.data.frame(explicit_model$item.factors)

Product_recom <- c()
for (i in Max){
  Product_recom <- c(Product_recom, dat$id[i])
}

df <- data.frame(users=1:20, Product_recom)

df %>%
  left_join(products, by=c("Product_recom" = "product_id"), copy=T) %>%
  select(users, Product_recom, product_name)
```

```
##    users Product_recom
## 1      1           196
## 2      2         32792
## 3      3         39190
## 4      4           196
```

```
## 5      5         26604
## 6      6         32792
## 7      7         40852
## 8      8         40852
## 9      9         27973
## 10    10         28535
## 11    11           196
## 12    12         32792
## 13    13         39190
## 14    14           196
## 15    15         26604
## 16    16         32792
## 17    17         40852
## 18    18         40852
## 19    19         27973
## 20    20         28535
##                                                    product_name
## 1                                                          Soda
## 2                                  Chipotle Beef & Pork Realstick
## 3                                Vanilla Unsweetened Almond Milk
## 4                                                          Soda
## 5                                            Organic Blackberries
## 6                                  Chipotle Beef & Pork Realstick
## 7                                      Lactose Free Fat Free Milk
## 8                                      Lactose Free Fat Free Milk
## 9  Almond Non-Dairy Yogurt Made From Real Almonds Plain Low Fat
## 10                                    Cucumber & Garlic  Tzatziki
## 11                                                         Soda
## 12                                 Chipotle Beef & Pork Realstick
## 13                               Vanilla Unsweetened Almond Milk
## 14                                                         Soda
## 15                                           Organic Blackberries
## 16                                 Chipotle Beef & Pork Realstick
## 17                                     Lactose Free Fat Free Milk
## 18                                     Lactose Free Fat Free Milk
## 19 Almond Non-Dairy Yogurt Made From Real Almonds Plain Low Fat
## 20                                    Cucumber & Garlic  Tzatziki
```

bartekskorulski@gmail.com