Instacart Exploratory Analysis

```
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
```

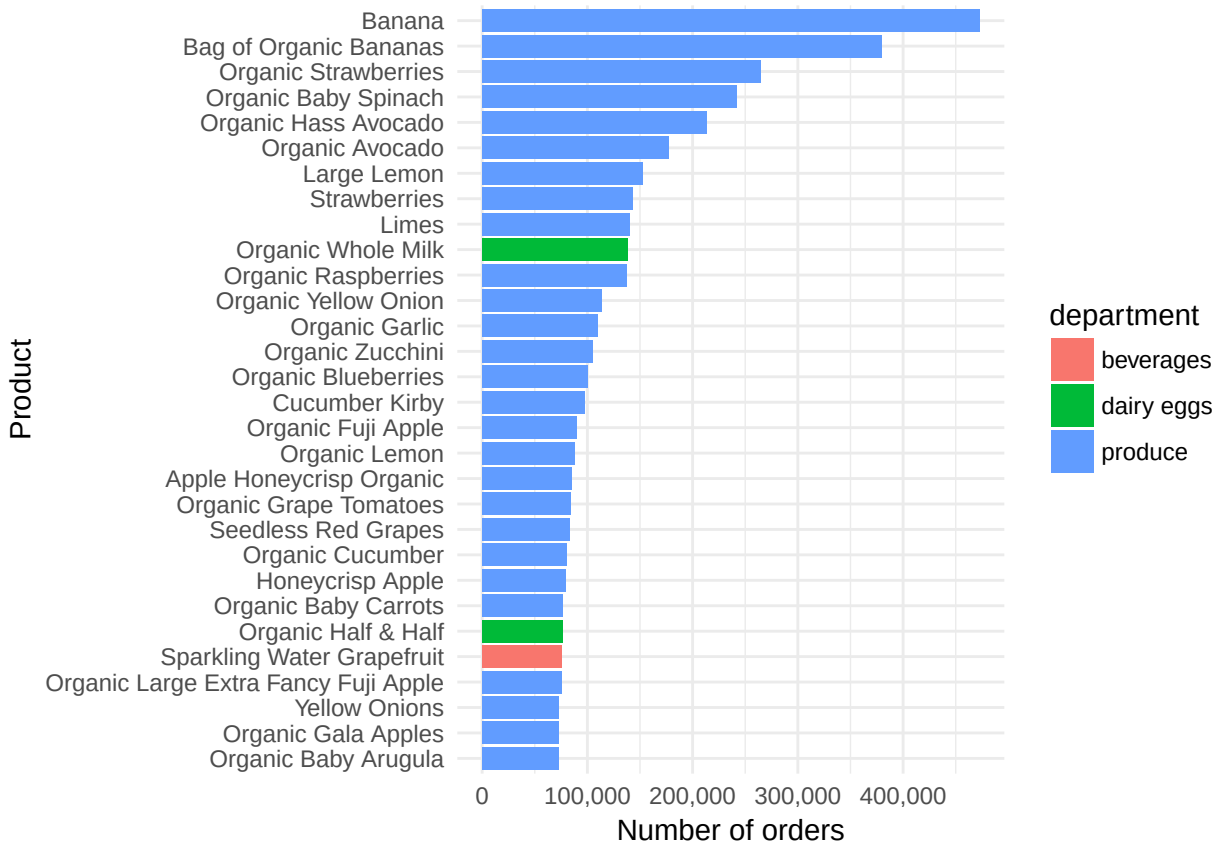
```
readInstacart()

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"
```

Most popular product

```
dbGetQuery(sc, "
SELECT product_id, product_name, n_orders, department
FROM (
    SELECT op.product_id, product_name, n_orders, department_id
    FROM (
        SELECT product_id, COUNT(1) AS n_orders
        FROM order_products__prior_tbl
        GROUP BY product_id
        ORDER BY n_orders DESC
        LIMIT 30
    ) op
    LEFT JOIN (
        SELECT product_id, product_name, department_id
        FROM products_tbl
    ) p
    ON op.product_id = p.product_id
) pp
LEFT JOIN (
    SELECT department_id, department
    FROM departments_tbl
) d
ON pp.department_id = d.department_id") %>%
```

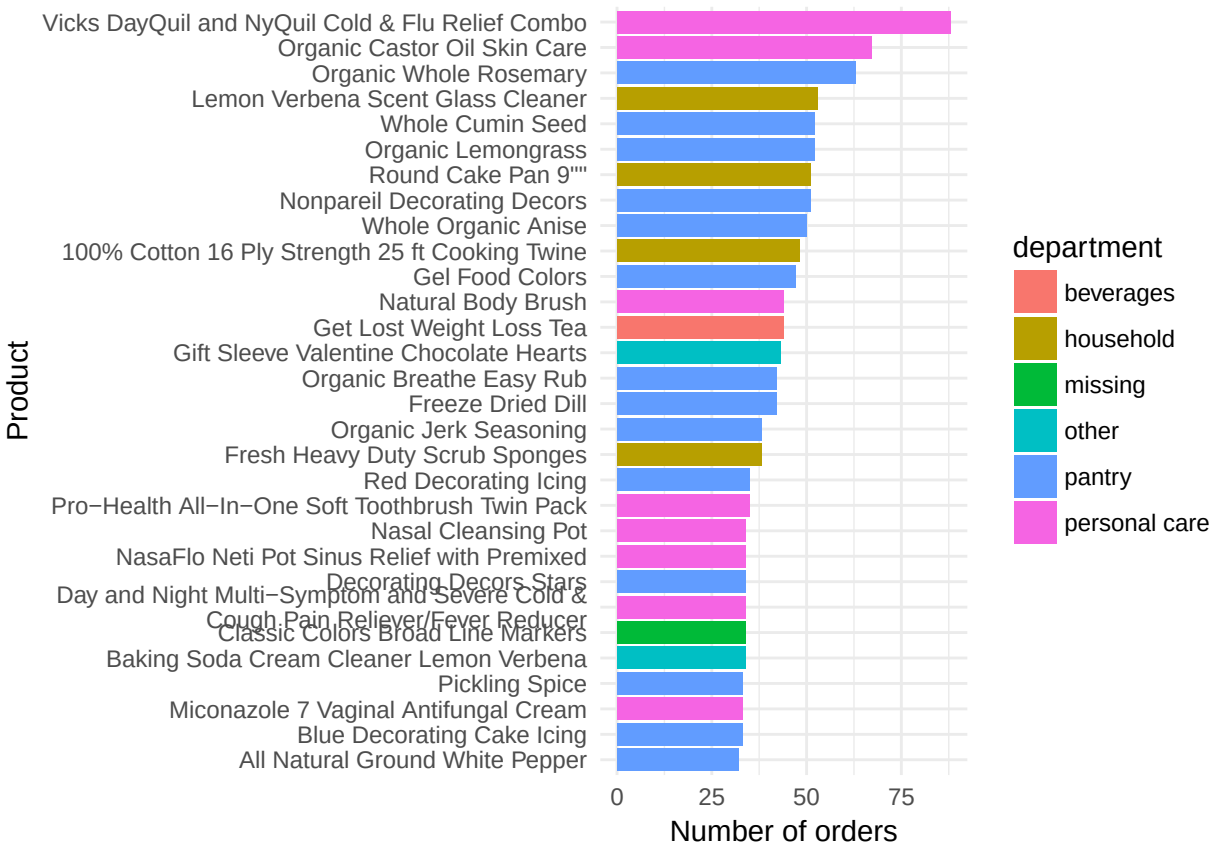
```
  ggplot(aes(x=reorder(product_name,n_orders),y=n_orders,fill=department)) +
  geom_bar(stat="identity") +
  scale_y_continuous(label=scales::comma) +
  coord_flip() +
  xlab("Product") +
  ylab("Number of orders") +
  theme_minimal()
```



Products never bought again

```
dbGetQuery(sc, "
SELECT product_id, product_name, n_orders, department
FROM (
    SELECT op.product_id, product_name, n_orders, department_id
    FROM (
        SELECT product_id, COUNT(1) AS n_orders
        FROM order_products__prior_tbl
        GROUP BY product_id
        HAVING SUM(reordered) = 0
        ORDER BY n_orders DESC
        LIMIT 30
    ) op
    LEFT JOIN (
        SELECT product_id, product_name, department_id
```

```
        FROM products_tbl
    ) p
    ON op.product_id = p.product_id
) pp
LEFT JOIN (
    SELECT department_id, department
    FROM departments_tbl
) d
ON pp.department_id = d.department_id") %>%
  ggplot(aes(x=reorder(stringr::str_wrap(product_name, 50),n_orders), y=n_orders, fill=department)) +
  geom_bar(stat="identity") +
  scale_y_continuous(label=scales::comma) +
  coord_flip() +
  xlab("Product") +
  ylab("Number of orders") +
  theme_minimal()
```



Most repeated first time bought product

```
dbGetQuery(sc, "
SELECT pp.product_id, product_name, COUNT(1) AS n_orders
FROM (
    SELECT o.order_id, product_id
    FROM (
```

```
      SELECT order_id
      FROM orders_tbl
      WHERE eval_set = 'prior' AND order_number = 1
    ) o
    LEFT JOIN (
        SELECT order_id, product_id
        FROM order_products__prior_tbl
    ) op
    ON o.order_id = op.order_id
) pp
LEFT JOIN (
    SELECT product_id, product_name, department_id
    FROM products_tbl
) p
ON pp.product_id = p.product_id
GROUP BY pp.product_id, product_name, department_id
ORDER BY n_orders DESC
LIMIT 10") %>%
  ggplot(aes(x=reorder(product_name,n_orders), y=n_orders)) +
  geom_bar(stat="identity") +
  scale_y_continuous(label=scales::comma) +
  coord_flip() +
  xlab("Product") +
  ylab("Number of first orders") +
  theme_minimal()
```