p8105_hw3_cz2955

2025-10-06

Problem 1

Load data:

```
library(p8105.datasets)
data("instacart")
```

This dataset comprises partial information from Instacart order data, containing users' historical orders and itemised details of goods within those orders. Each row in the data represents a specific item purchased by a user during a particular order. Key variables include order_id, product_id, add_to_cart_order, reordered, user_id, eval_set, order_number, order_dow, order_hour_of_day, days_since_prior_order, product_name, aisle_id, department_id, aisle, department. There are total 1384617variables. The average reorder rate across all items is 0.599, indicating that a significant fraction of purchases are repeat orders. ### (a):

```
length(unique(instacart$aisle))
```

```
## [1] 134
```

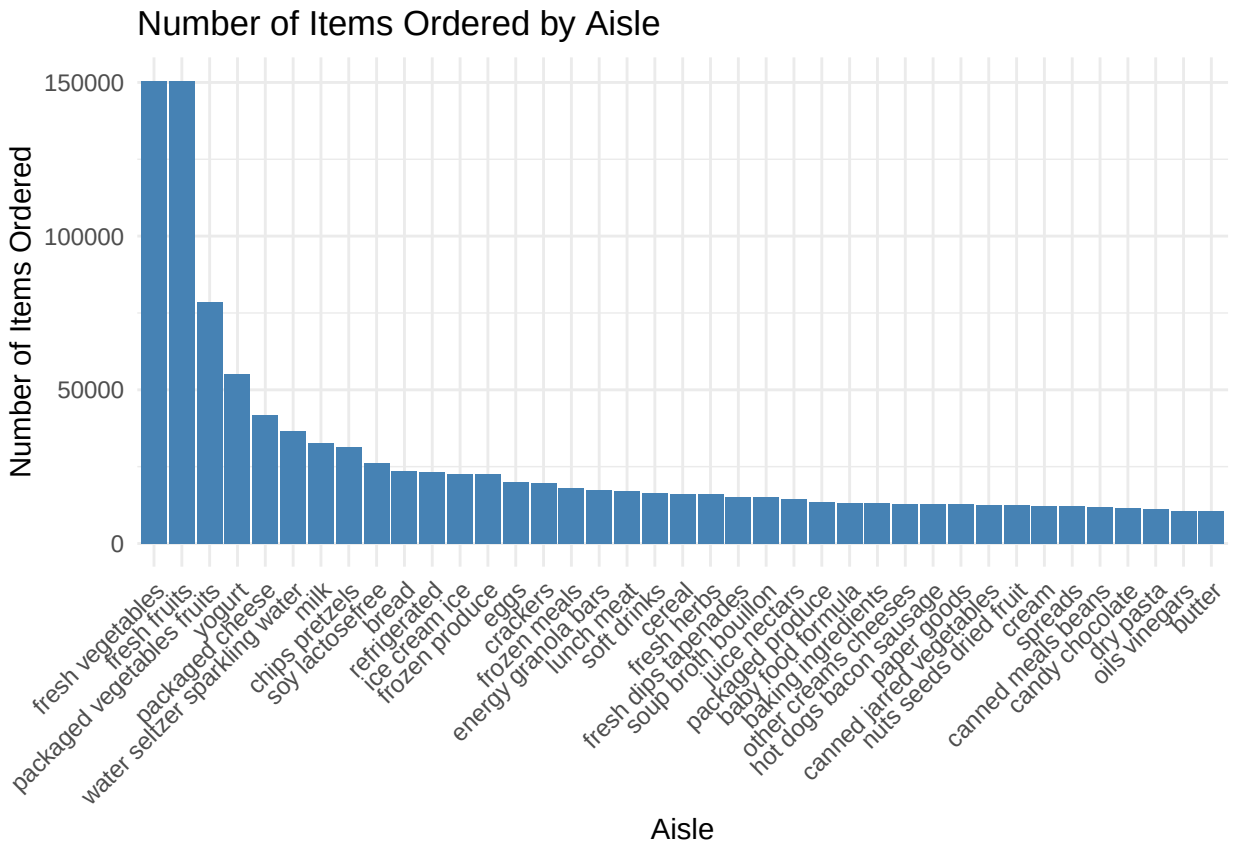
```
aisle_counts <- instacart%>%
  group_by(aisle)%>%
  summarise(num_orders = n())%>%
  arrange(desc(num_orders))

head(aisle_counts)
```

```
## # A tibble: 6 x 2
##   aisle                         num_orders
##   <chr>                              <int>
## 1 fresh vegetables                  150609
## 2 fresh fruits                      150473
## 3 packaged vegetables fruits         78493
## 4 yogurt                             55240
## 5 packaged cheese                    41699
## 6 water seltzer sparkling water      36617
```

(b):

```
aisle_counts_large <- aisle_counts %>%
  filter(num_orders > 10000)
ggplot(aisle_counts_large, aes(x = reorder(aisle, -num_orders), y = num_orders)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Number of Items Ordered by Aisle",
       x = "Aisle",
       y = "Number of Items Ordered") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



(c):

```
popular_items <- instacart %>%
  filter(aisle %in% c("baking ingredients", "dog food care", "packaged vegetables fruits")) %>%
  group_by(aisle, product_name) %>%
  summarise(num_orders = n()) %>%
  arrange(aisle, desc(num_orders)) %>%
  slice_head(n = 3)
```

```
## `summarise()` has grouped output by 'aisle'. You can override using the
## `.groups` argument.
```

```
popular_items
```

```
## # A tibble: 9 x 3
## # Groups:   aisle [3]
##   aisle                      product_name                                num_orders
##   <chr>                      <chr>                                            <int>
## 1 baking ingredients         Light Brown Sugar                                  499
## 2 baking ingredients         Pure Baking Soda                                   387
## 3 baking ingredients         Cane Sugar                                         336
## 4 dog food care              Snack Sticks Chicken & Rice Recipe Dog ~            30
## 5 dog food care              Organix Chicken & Brown Rice Recipe                28
## 6 dog food care              Small Dog Biscuits                                 26
## 7 packaged vegetables fruits Organic Baby Spinach                             9784
## 8 packaged vegetables fruits Organic Raspberries                              5546
## 9 packaged vegetables fruits Organic Blueberries                              4966
```

(d)

```
mean_order_hour <- instacart %>%
  filter(product_name %in% c("Pink Lady Apples", "Coffee Ice Cream")) %>%
  group_by(product_name, order_dow) %>%
  summarise(mean_hour = mean(order_hour_of_day, na.rm = TRUE)) %>%
  pivot_wider(names_from = order_dow, values_from = mean_hour) %>%
  arrange(product_name)
```

```
## `summarise()` has grouped output by 'product_name'. You can override using the
## `.groups` argument.
```

```
mean_order_hour
```

```
## # A tibble: 2 x 8
## # Groups:   product_name [2]
##   product_name       `0`   `1`   `2`   `3`   `4`   `5`   `6`
##   <chr>            <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Coffee Ice Cream  13.8  14.3  15.4  15.3  15.2  12.3  13.8
## 2 Pink Lady Apples  13.4  11.4  11.7  14.2  11.6  12.8  11.9
```