UW 4th Datathon: Australian Grocery Store (Analytics and Visualization Part)

Zilin Huang, Vivian Han, Albert Hutiancong Wang, Daniel Luo

2023-02-19

Retrieve and clean the data:

```r
# Retrieve the data
setwd("/Users/huang/OneDrive/Desktop/CFRM 425")
sales = read.csv("sales_data_2017_2018_for_tableau_with_new_date_columns.csv")

# Clearn the data by filling all null values with the mean of corresponding column:
for(i in 1:ncol(sales)){
  sales[is.na(sales[,i]), i] <- mean(sales[,i], na.rm = TRUE)
}
```

```
## Warning in mean.default(sales[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(sales[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(sales[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(sales[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(sales[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(sales[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(sales[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(sales[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA
```

Analyze:

We choose this question: Which items should the store stop selling? Why?

Answer:

For items that the store should stop selling, they must be low in number of purchase and/or low in total profits generated. For simplicity, we will only consider the former case.

We analyze this case by past years (2017 and 2018) by sub-setting the original sales' data into two parts:

```r
# Subset the sales data by two years
sales_17 <- sales[sales["year"] == 2017, ]
sales_18 <- sales[sales["year"] == 2018, ]
```

Then, we count the number of receipts for all of the items, as an index that implies the ability to generate profits for each of them:

```r
# Group the data by number of receipents ID's for each item:
agg_17 <- aggregate(sales_17$receipt_id, by=list(sales_17$item_name), FUN=length)
agg_17 <- agg_17[agg_17$x <= 10, ]

agg_18 <- aggregate(sales_18$receipt_id, by=list(sales_18$item_name), FUN=length)
agg_18 <- agg_18[agg_18$x <= 10, ]
```

(One point to notice is that it is not recommended to analyze quantity, since it tends to be influenced by other factors like size or weight of the product. For instance, Dragonfruits has less quantity sold than Kiwi, but higher price per unit)

Visual:

More details in the slides.