Data Cleaning and Preparation

library(readr)  
library(dplyr)  
library(tidyr)  
library(arules)  
library(arulesViz)  
library(methods)  
library(data.table)  
library(magrittr)  
library(knitr)  
library(ggplot2)   
library(treemap)  
library(wordcloud)  
library(stringr)

# Step 0: Load the data in RStudio

Aisles <- fread('aisles.csv')  
Departments <- fread('departments.csv')  
Products <- fread("products.csv")  
Orders <- fread('orders.csv')  
Order\_Products\_Prior <- fread("order\_products\_\_prior.csv")  
Order\_Products\_Train <- fread('order\_products\_\_train.csv')

# Step 1: Glimpse of the data

**The Dim function provides the number of obervations(rows) and variables(columns) for each table.**

dim(Aisles)

## [1] 134 2

dim(Departments)

## [1] 21 2

dim(Products)

## [1] 49688 4

dim(Orders)

## [1] 3421083 7

dim(Order\_Products\_Prior)

## [1] 32434489 4

**The Name function provides the column names of each table.**

names(Aisles)

## [1] "aisle\_id" "aisle"

names(Departments)

## [1] "department\_id" "department"

names(Products)

## [1] "product\_id" "product\_name" "aisle\_id" "department\_id"

names(Orders)

## [1] "order\_id" "user\_id"   
## [3] "eval\_set" "order\_number"   
## [5] "order\_dow" "order\_hour\_of\_day"   
## [7] "days\_since\_prior\_order"

names(Order\_Products\_Prior)

## [1] "order\_id" "product\_id" "add\_to\_cart\_order"  
## [4] "reordered"

**The Glimpse function provides the structure of each table.**

glimpse(Aisles)

## Observations: 134  
## Variables: 2  
## $ aisle\_id <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...  
## $ aisle <chr> "prepared soups salads", "specialty cheeses", "energy...

glimpse(Departments)

## Observations: 21  
## Variables: 2  
## $ department\_id <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...  
## $ department <chr> "frozen", "other", "bakery", "produce", "alcohol...

glimpse(Products)

## Observations: 49,688  
## Variables: 4  
## $ product\_id <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...  
## $ product\_name <chr> "Chocolate Sandwich Cookies", "All-Seasons Salt"...  
## $ aisle\_id <int> 61, 104, 94, 38, 5, 11, 98, 116, 120, 115, 31, 1...  
## $ department\_id <int> 19, 13, 7, 1, 13, 11, 7, 1, 16, 7, 7, 1, 11, 17,...

glimpse(Orders)

## Observations: 3,421,083  
## Variables: 7  
## $ order\_id <int> 2539329, 2398795, 473747, 2254736, 4315...  
## $ user\_id <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, ...  
## $ eval\_set <chr> "prior", "prior", "prior", "prior", "pr...  
## $ order\_number <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 1, 2...  
## $ order\_dow <int> 2, 3, 3, 4, 4, 2, 1, 1, 1, 4, 4, 2, 5, ...  
## $ order\_hour\_of\_day <int> 8, 7, 12, 7, 15, 7, 9, 14, 16, 8, 8, 11...  
## $ days\_since\_prior\_order <dbl> NA, 15, 21, 29, 28, 19, 20, 14, 0, 30, ...

glimpse(Order\_Products\_Prior)

## Observations: 32,434,489  
## Variables: 4  
## $ order\_id <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3,...  
## $ product\_id <int> 33120, 28985, 9327, 45918, 30035, 17794, 401...  
## $ add\_to\_cart\_order <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 2, 3, 4, 5, 6,...  
## $ reordered <int> 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,...

**The Summary function provides the descriptive statistics of each table.**

summary(Aisles)

## aisle\_id aisle   
## Min. : 1.00 Length:134   
## 1st Qu.: 34.25 Class :character   
## Median : 67.50 Mode :character   
## Mean : 67.50   
## 3rd Qu.:100.75   
## Max. :134.00

summary(Departments)

## department\_id department   
## Min. : 1 Length:21   
## 1st Qu.: 6 Class :character   
## Median :11 Mode :character   
## Mean :11   
## 3rd Qu.:16   
## Max. :21

summary(Products)

## product\_id product\_name aisle\_id department\_id   
## Min. : 1 Length:49688 Min. : 1.00 Min. : 1.00   
## 1st Qu.:12423 Class :character 1st Qu.: 35.00 1st Qu.: 7.00   
## Median :24844 Mode :character Median : 69.00 Median :13.00   
## Mean :24844 Mean : 67.77 Mean :11.73   
## 3rd Qu.:37266 3rd Qu.:100.00 3rd Qu.:17.00   
## Max. :49688 Max. :134.00 Max. :21.00

summary(Orders)

## order\_id user\_id eval\_set order\_number   
## Min. : 1 Min. : 1 Length:3421083 Min. : 1.00   
## 1st Qu.: 855272 1st Qu.: 51394 Class :character 1st Qu.: 5.00   
## Median :1710542 Median :102689 Mode :character Median : 11.00   
## Mean :1710542 Mean :102978 Mean : 17.15   
## 3rd Qu.:2565812 3rd Qu.:154385 3rd Qu.: 23.00   
## Max. :3421083 Max. :206209 Max. :100.00   
##   
## order\_dow order\_hour\_of\_day days\_since\_prior\_order  
## Min. :0.000 Min. : 0.00 Min. : 0.00   
## 1st Qu.:1.000 1st Qu.:10.00 1st Qu.: 4.00   
## Median :3.000 Median :13.00 Median : 7.00   
## Mean :2.776 Mean :13.45 Mean :11.11   
## 3rd Qu.:5.000 3rd Qu.:16.00 3rd Qu.:15.00   
## Max. :6.000 Max. :23.00 Max. :30.00   
## NA's :206209

summary(Order\_Products\_Prior)

## order\_id product\_id add\_to\_cart\_order reordered   
## Min. : 2 Min. : 1 Min. : 1.000 Min. :0.0000   
## 1st Qu.: 855943 1st Qu.:13530 1st Qu.: 3.000 1st Qu.:0.0000   
## Median :1711048 Median :25256 Median : 6.000 Median :1.0000   
## Mean :1710749 Mean :25576 Mean : 8.351 Mean :0.5897   
## 3rd Qu.:2565514 3rd Qu.:37935 3rd Qu.: 11.000 3rd Qu.:1.0000   
## Max. :3421083 Max. :49688 Max. :145.000 Max. :1.0000

# Step 2: Checking for any missing values in each table column wise

**Aisles Dataset**

data.frame(missing.values = sapply(Aisles, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## aisle\_id 0  
## aisle 0

**Departments Dataset**

data.frame(missing.values = sapply(Departments, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## department\_id 0  
## department 0

**Products Dataset**

data.frame(missing.values = sapply(Products, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## product\_id 0  
## product\_name 0  
## aisle\_id 0  
## department\_id 0

\*\* Orders Dataset\*\*

data.frame(missing.values = sapply(Orders, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## order\_id 0  
## user\_id 0  
## eval\_set 0  
## order\_number 0  
## order\_dow 0  
## order\_hour\_of\_day 0  
## days\_since\_prior\_order 206209

**Order\_Products\_Prior Dataset**

data.frame(missing.values = sapply(Order\_Products\_Prior, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## order\_id 0  
## product\_id 0  
## add\_to\_cart\_order 0  
## reordered 0

**Order\_Products\_Train Dataset**

data.frame(missing.values = sapply(Order\_Products\_Train, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## order\_id 0  
## product\_id 0  
## add\_to\_cart\_order 0  
## reordered 0

# Step 3: Data Preparation

**Converting Continuous variable to Categorical for better interpretation**

Orders<-Orders %>% mutate(eval\_set=as.factor(eval\_set))  
Products<-Products %>% mutate(product\_name=as.factor(product\_name))  
Aisles <- Aisles %>% mutate(aisle=as.factor(aisle))  
Departments <- Departments %>% mutate(department=as.factor(department))