7313 Hand-in

Group 3

2021-11-22

```
#Creating connection to MySQL database
con = dbConnect(MySQL(), dbname = "BnL",
                host = "mysql-1.cda.hhs.se", port = 3306,
                user = "bnl", password = "bnl@sse")
```

```
#Loading data into R
df <- dbGetQuery(con,"SELECT item, returned, amount, discount, quantity, dept_desc, sustainability_id_d
                     FROM Transactions t LEFT JOIN Products p
                     USING(item)")
glimpse(df)
```

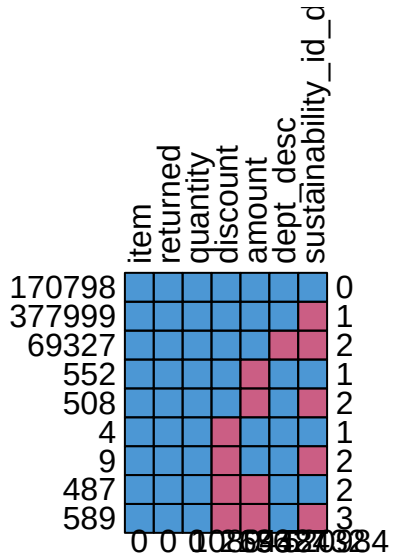
```
## Rows: 620,273
## Columns: 7
## $ item                   <dbl> 43708023, 16976458, 42641555, 43573146, 4388504~
## $ returned               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ amount                 <int> 0, 1800, 1800, 880, 2960, 0, 0, 0, 0, 3300, 0, ~
## $ discount               <int> 0, 0, 0, 220, 740, 0, 0, 0, 0, 0, 0, 0, 744, 0,~
## $ quantity               <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1,~
## $ dept_desc              <chr> "Receipt texts", "Makeup B", "Makeup B", "Hair ~
## $ sustainability_id_desc <chr> NA, NA, NA, "Non sustainable", NA, NA, NA, NA, ~
```

```
#summary(df)
```

```
#Checking for missing values
contains.missing <- df %>%
  filter_all(any_vars(is.na(.))) %>%
  select_if(function(x) any(is.na(x)))
#contains.missing

df %>%
  sapply(function(x) sum(is.na(x)))
```

```
##                    item                returned                  amount
##                       0                       0                    2136
##                discount                quantity               dept_desc
##                    1089                       0                   69327
## sustainability_id_desc
##                  448432
```

```
md.pattern(df, rotate.names = T)
```




```
##        item returned quantity discount amount dept_desc sustainability_id_desc
## 170798    1        1        1        1      1         1                       1
## 377999    1        1        1        1      1         1                       0
## 69327     1        1        1        1      1         0                       0
## 552       1        1        1        1      0         1                       1
## 508       1        1        1        1      0         1                       0
## 4         1        1        1        0      1         1                       1
## 9         1        1        1        0      1         1                       0
## 487       1        1        1        0      0         1                       1
## 589       1        1        1        0      0         1                       0
##           0        0        0     1089   2136     69327                  448432
##
## 170798      0
## 377999      1
## 69327       2
## 552         1
## 508         2
## 4           1
## 9           2
## 487         2
## 589         3
##        520984
```

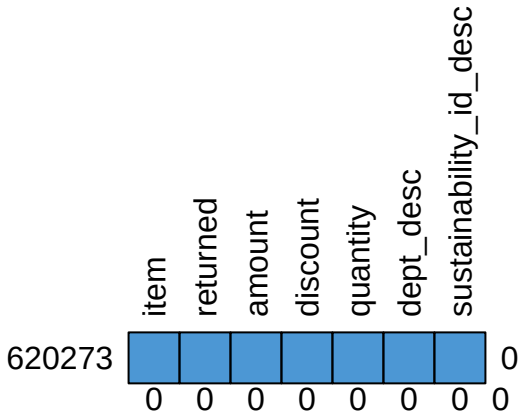
```
#Imputation of missing values
##Mode imputation for variables of type character and mean imputation for variables of type double
mode_dept_desc <- df %>%
  filter(!is.na(dept_desc)) %>%
  count(dept_desc) %>%
  top_n(1, n) %>%
  select(dept_desc) %>%
  unlist(use.names = F)
mode_sustainability_id_desc <- df %>%
```

```
  filter(!is.na(sustainability_id_desc)) %>%
  count(sustainability_id_desc) %>%
  top_n(1, n) %>%
  select(sustainability_id_desc) %>%
  unlist(use.names = F)

df_clean <- df %>%
  mutate(amount = ifelse(is.na(amount), mean(amount, na.rm = T), amount),
         discount = ifelse(is.na(discount), mean(discount, na.rm = T), discount),
         dept_desc = ifelse(is.na(dept_desc), mode_dept_desc, dept_desc),
         sustainability_id_desc = ifelse(is.na(sustainability_id_desc), mode_sustainability_id_desc, su

#Final check
df_clean %>%
  sapply(function(x) sum(is.na(x)))
```

```
##                    item                returned                  amount
##                       0                       0                       0
##                discount                quantity               dept_desc
##                       0                       0                       0
## sustainability_id_desc
##                       0
```

```
md.pattern(df_clean, rotate.names = T)
```

```
##  /\     /\
## {  `---'  }
## {  0   0  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```




```
##        item returned amount discount quantity dept_desc sustainability_id_desc
```

```
## 620273    1        1      1        1        1        1                    1
##           0        0      0        0        0        0                    0
##
## 620273 0
##        0
```

```
#summary(df_clean)
glimpse(df_clean)
```

```
## Rows: 620,273
## Columns: 7
## $ item                   <dbl> 43708023, 16976458, 42641555, 43573146, 4388504~
## $ returned               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ amount                 <dbl> 0, 1800, 1800, 880, 2960, 0, 0, 0, 0, 3300, 0, ~
## $ discount               <dbl> 0, 0, 0, 220, 740, 0, 0, 0, 0, 0, 0, 0, 744, 0,~
## $ quantity               <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1,~
## $ dept_desc              <chr> "Receipt texts", "Makeup B", "Makeup B", "Hair ~
## $ sustainability_id_desc <chr> "Non sustainable", "Non sustainable", "Non sust~
```

```
#Remove modes + support table
remove(mode_dept_desc, mode_sustainability_id_desc, contains.missing)
```

```
#Exclusion of non-product-related observations (Receipt texts, Gift With Purchase, Marketing Material, 
df_clean <- df_clean %>%
  mutate(no_product = ifelse(dept_desc == "Receipt texts", 1,
                             ifelse(dept_desc == "Gift With Purchase", 1,
                                    ifelse(dept_desc == "Marketing Material", 1,
                                           ifelse(dept_desc == "Sales Kicks E-commerce", 1, 0)))))
df_clean <- df_clean %>%
  filter(no_product == 0)
df_clean <- df_clean %>%
  select(-no_product)
```

```
#Creation of new variables
##Sustainable
df_clean <- df_clean %>%
  mutate(sustainable = ifelse(sustainability_id_desc == "Environmentally labelled", "sustainable",
                              ifelse(sustainability_id_desc == "Organic", "sustainable",
                                     ifelse(sustainability_id_desc == "Social responsibility", "sustaina
                                            ifelse(sustainability_id_desc == "Sustainable material", "su

##Category
df_clean  <- df_clean  %>%
  mutate(category = ifelse(dept_desc == "Makeup B", "Makeup",
                           ifelse(dept_desc == "Makeup PL", "Makeup",
                                  ifelse(dept_desc == "Face B", "Face",
                                         ifelse(dept_desc == "Face PL", "Face",
                                                ifelse(dept_desc == "Body B", "Body",
                                                       ifelse(dept_desc == "Body PL", "Body",
                                                              ifelse(dept_desc == "Hair B", "Hair",
                                                                     ifelse(dept_desc == "Hair PL", "Ha
                                                                            ifelse(dept_desc == "Hair 
```

```
                                                                    ifelse(dept_desc ==
                                                                           ifelse(dept_
                                                                                  ifels
```

```
#Final check
glimpse(df_clean)
```

```
## Rows: 523,851
## Columns: 9
## $ item                   <dbl> 16976458, 42641555, 43573146, 43885045, 4432024~
## $ returned               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ amount                 <dbl> 1800, 1800, 880, 2960, 0, 3300, 0, 5200, 946, 3~
## $ discount               <dbl> 0, 0, 220, 740, 0, 0, 0, 0, 744, 0, 0, 0, 0, 0,~
## $ quantity               <int> 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1,~
## $ dept_desc              <chr> "Makeup B", "Makeup B", "Hair B", "Face B", "Ma~
## $ sustainability_id_desc <chr> "Non sustainable", "Non sustainable", "Non sust~
## $ sustainable            <chr> "non-sustainable", "non-sustainable", "non-sust~
## $ category               <chr> "Makeup", "Makeup", "Hair", "Face", "Makeup", "~
```

```
df_clean %>%
  group_by(category, dept_desc) %>%
  count()
```

```
## # A tibble: 21 x 3
## # Groups:   category, dept_desc [21]
##    category         dept_desc                n
##    <chr>            <chr>                <int>
##  1 Accessories      Childrens Acc C          1
##  2 Accessories      Womens Accessories C  5851
##  3 Accessories      Womens Bags PL           2
##  4 Body             Body B               35097
##  5 Body             Body PL              22155
##  6 Childrens Care   Childrens Care B        93
##  7 Cooking & Dining Cooking & Dining B      64
##  8 Face             Face B               68973
##  9 Face             Face PL               5746
## 10 Fragrance        Fragrance B          27505
## # ... with 11 more rows
```

```
#Removal of 'old' variables
df_clean <- df_clean %>%
  select(-dept_desc, -sustainability_id_desc)
```

```r
#Ensuring correct aggregation levels of variables
df_final <- df_clean %>%
  group_by(item) %>%
  summarize(total_quantity = sum(quantity[quantity>=0]),
            avg_amount = mean(amount[amount>=0]),
            avg_discount = mean(discount[discount>=0])/mean(amount[amount>=0]),
            per_returned = sum(abs(quantity)[returned==1])/sum(quantity[quantity>=0]),
            mode_category = mfv(category),
            mode_sustainable = mfv(sustainable))
```

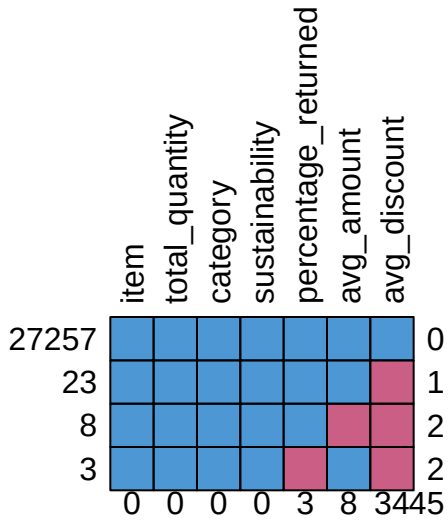
```r
#Renaming of columns/ variable
df_final <- df_final %>%
  rename(category = mode_category,
         sustainability = mode_sustainable,
         percentage_returned = per_returned)

#Ensuring correct data types
glimpse(df_final)
```

```
## Rows: 27,291
## Columns: 7
## $ item                <dbl> 275040, 275057, 276584, 1264647, 1264654, 1266147,~
## $ total_quantity      <int> 200, 206, 1, 10, 26, 78, 3, 2, 22, 32, 1, 3, 101, ~
## $ avg_amount          <dbl> 648.5185, 651.7767, 150.0000, 1650.0000, 1568.8462~
## $ avg_discount        <dbl> 0.01714460, 0.02071262, 0.00000000, 0.00000000, 0.~
## $ percentage_returned <dbl> 0.005000000, 0.009708738, 1.000000000, 0.000000000~
## $ category            <chr> "Makeup", "Makeup", "Face", "Face", "Face", "Body"~
## $ sustainability      <chr> "non-sustainable", "non-sustainable", "non-sustain~
```

```r
df_final <- df_final %>%
  mutate(item = as.factor(item),
         sustainability = as.factor(sustainability),
         category = as.factor(category))

#Checking for missing values
df_final %>%
  sapply(function(x) sum(is.na(x)))
```

```
##                item      total_quantity           avg_amount        avg_discount
##                   0                   0                   8                  34
## percentage_returned            category      sustainability
##                   3                   0                   0
```

```r
md.pattern(df_final, rotate.names = T)
```




```
##       item total_quantity category sustainability percentage_returned
## 27257    1              1        1              1                    1
## 23       1              1        1              1                    1
## 8        1              1        1              1                    1
## 3        1              1        1              1                    0
##          0              0        0              0                    3
##       avg_amount avg_discount
## 27257          1            1  0
## 23             1            0  1
## 8              0            0  2
## 3              1            0  2
##                8           34 45
```

```
#Deleting 34 items w/ missing values - missing values arising due to failures in calculations for aggre
df_final <- df_final %>%
    filter_all(all_vars(!is.na(.)))
```

```
#Final presentation of variables
head(df_final, 10)
```

```
## # A tibble: 10 x 7
##    item    total_quantity avg_amount avg_discount percentage_returned category
##    <fct>            <int>      <dbl>        <dbl>               <dbl> <fct>
##  1 275040             200       649.       0.0171             0.005   Makeup
##  2 275057             206       652.       0.0207             0.00971 Makeup
##  3 276584               1       150        0                  1       Face
##  4 1264647             10      1650        0                  0       Face
##  5 1264654             26      1569.       0.0579             0       Face
##  6 1266147             78      1727.       0.0284             0       Body
##  7 1267194              3      2250        0                  0       Fragrance
##  8 1269471              2      2674        0.522              0       Fragrance
##  9 1279181             22      5632.       0.149              0.0455  Face
## 10 1279595             32      2778.       0.0427             0.0312  Face
## # ... with 1 more variable: sustainability <fct>
```

```r
glimpse(df_final)
```

```
## Rows: 27,257
## Columns: 7
## $ item                <fct> 275040, 275057, 276584, 1264647, 1264654, 1266147,~
## $ total_quantity      <int> 200, 206, 1, 10, 26, 78, 3, 2, 22, 32, 1, 3, 101, ~
## $ avg_amount          <dbl> 648.5185, 651.7767, 150.0000, 1650.0000, 1568.8462~
## $ avg_discount        <dbl> 0.01714460, 0.02071262, 0.00000000, 0.00000000, 0.~
## $ percentage_returned <dbl> 0.005000000, 0.009708738, 1.000000000, 0.000000000~
## $ category            <fct> Makeup, Makeup, Face, Face, Face, Body, Fragrance,~
## $ sustainability      <fct> non-sustainable, non-sustainable, non-sustainable,~
```

```r
#aggregation level: item
#target: percentage_returned
#predictors: total_quantity, avg_amount, avg_discount, category, sustainability
```