R\_final\_project

Christopher Zheng - 260760794

19/11/2019

# ——————————————————————

# Preliminaries

##### To begin with, we load libraries that are necessary for the project and import the Women’s Clothing Dataset. Note that we only select six fields relevant for analysis, i.e., Review\_ID, Clothing\_ID, Age, Rating, Recommended, and Department\_Name.

# load dependencies  
library(ggplot2)  
library(dplyr)  
library(tidyverse)  
library(forcats)  
library(readr)  
  
# Read data and select fields  
Womens\_Clothing\_Reviews <- read\_csv("D:/PERSONAL/McGill/McGill/McGill Current/MATH 208/assignments/R\_final\_project/Womens\_Clothing\_Reviews.csv")  
data <- Womens\_Clothing\_Reviews[c("Review\_ID","Clothing\_ID","Age","Rating","Recommended","Department\_Name")]

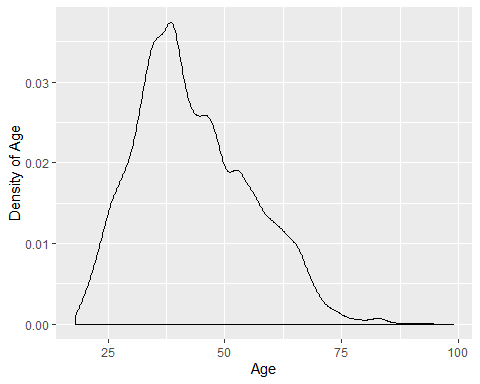
# ——————————————————————

# Task 1

### Summary of “Age”

##### The “Age” denotes the age of the reviewer in an integer. This variable spans from 18 to 99 and has a relatively right-skewed distribution as shown by the density plot below. A large proportion of the age values fall between 35 and 45 containing the mean (43.2) and median (41), and this age interval represents the group of people who are into leaving product reviews the most.

# Plot  
ggplot(data,aes(x=Age)) + geom\_density() +  
 xlab("Age") + ylab("Density of Age")



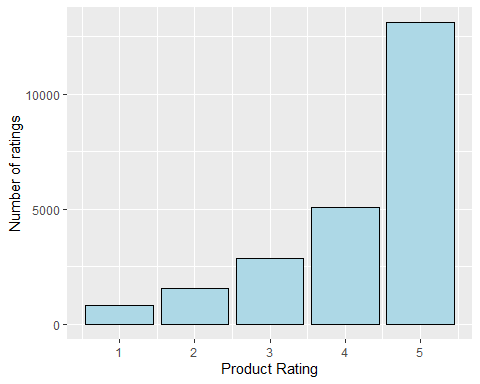
# Summary table  
data %>% summarise(Avg = mean(Age),  
 Med = median(Age),  
 '25%ile' = quantile(Age,0.25),  
 '75%ile' = quantile(Age,0.75),  
 StD = sd(Age),  
 IQR = IQR(Age)  
 )

## # A tibble: 1 x 6  
## Avg Med `25%ile` `75%ile` StD IQR  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 43.2 41 34 52 12.3 18

### Summary of “Rating”

##### The “Rating” variable denotes the integer rating a purchaser gives to a product and takes values from 1, 2, 3, 4, and 5, where 1 is for the worst and 5 is for the best. The distribution of Rating is heavily skewed towards the right, i.e., good ratings (5 and 4). This can be seen from the fact that the mean is 4.2 and the median is straight 5. The rating of 5 dominates with approximately 13,000 ratings, followed by about 5,000 ratings of 4. As the rating level decreases, the number of that specific rating decreases.

# plot  
ggplot(data,aes(x=Rating)) + geom\_bar(col='black',fill="lightblue") +  
 xlab("Product Rating") + ylab("Number of ratings")



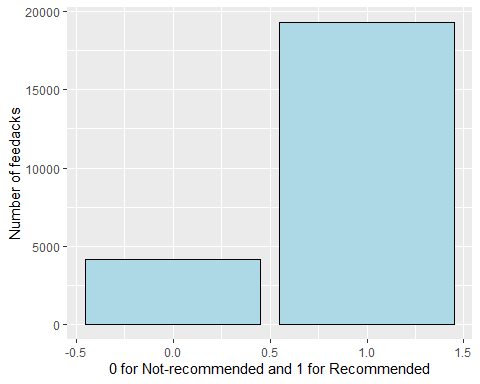
# Summary table  
data %>% summarise(Avg = mean(Rating),  
 Med = median(Rating),  
 '25%ile' = quantile(Rating,0.25),  
 '75%ile' = quantile(Rating,0.75),  
 StD = sd(Age),  
 IQR = IQR(Age)  
 )

## # A tibble: 1 x 6  
## Avg Med `25%ile` `75%ile` StD IQR  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 4.20 5 4 5 12.3 18

### Summary of “Recommended”

##### The “Recommended” variable is binary (0 or 1) and represents whether a purchaser is willing to recommend (denoted by 1) this or not (denoted by 0). The reviews with recommendation are 19,300 in number and those without recommendation are 4172 in number. Apparently, most reviews (82.22%) are with recommendations.

# plot  
ggplot(data,aes(x=Recommended)) + geom\_bar(col='black',fill="lightblue") +  
 xlab("0 for Not-recommended and 1 for Recommended") + ylab("Number of feedacks")



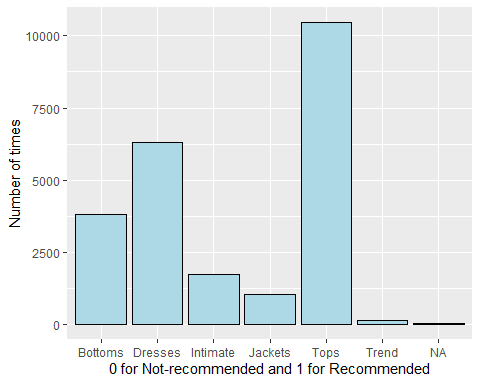
# helper  
counter <- function(x){  
 a <- table(data$Recommended)  
 b <- a[names(a)==x]  
 return(b[1])  
}  
  
# Summary table  
data %>% summarise(Avg = mean(Recommended),  
 Med = median(Recommended),  
 '25%ile' = quantile(Recommended,0.25),  
 '75%ile' = quantile(Recommended,0.75),  
 StD = sd(Recommended),  
 IQR = IQR(Recommended),  
 NotRecommended = counter(0),  
 Recommended = counter(1)  
 )

## # A tibble: 1 x 8  
## Avg Med `25%ile` `75%ile` StD IQR NotRecommended Recommended  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int> <int>  
## 1 0.822 1 1 1 0.382 0 4172 19314

### Summary of “Department\_Name”

##### The variable of “Department\_Name” denotes the name of the department that a product belongs to. The departments are Bottoms (3799 product reviews), Dresses (6319 product reviews), Intimate (1735 product reviews), Jackets (1032 product reviews), Tops (10468 product reviews), and Trend (119 product reviews). There are also 14 reviews associated with no department names, i.e., missing values. Tops and dresses are the two departments that house most reviews.

# plot  
ggplot(data,aes(x=Department\_Name)) + geom\_bar(col="black",fill="lightblue") +  
 xlab("0 for Not-recommended and 1 for Recommended") + ylab("Number of times")



counter <- function(x){  
 a <- table(data$Department\_Name)  
 b <- a[names(a)==x]  
 return(b[[x]])  
}  
  
  
# Summary table  
data %>% summarise(Bottoms = counter("Bottoms"),  
 Dresses = counter("Dresses"),  
 Intimate = counter("Intimate"),  
 Jackets = counter("Jackets"),  
 Tops = counter("Tops"),  
 Trend = counter("Trend"),  
 Missing = length(data$Department\_Name) - Bottoms - Dresses - Intimate - Jackets - Tops - Trend)

## # A tibble: 1 x 7  
## Bottoms Dresses Intimate Jackets Tops Trend Missing  
## <int> <int> <int> <int> <int> <int> <int>  
## 1 3799 6319 1735 1032 10468 119 14

### Remove data entries that miss values

##### As previously illustrated, Department\_Name contains 14 missing values which are supposed to be removed for further analysis.

# Remove rows whose department names are not the listed six valid departments.  
data <- data[!is.na(data$Department\_Name),]

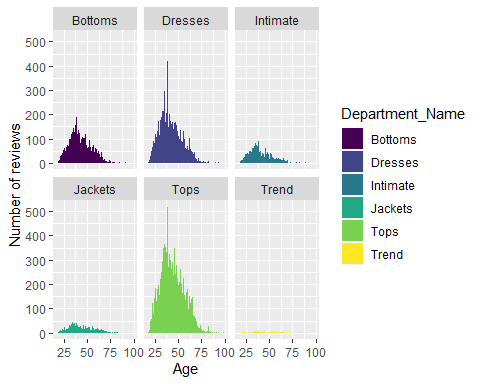
# ——————————————————————

# Task 2

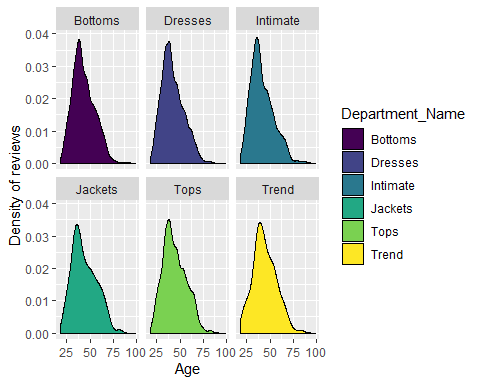
### Q1

##### If we separate the six departments apart w.r.t. to Age as in the following diagrams, we can tell that the six departments’ products are not equally popular under one scale in the barplot: Tops and Dresses received the largest number of reviews, whereas Trend the smallest, which is almost unnoticeable. However, if we graph the same data in a set of density plots, we discover that the distributions of the six departments’ product reviews are similar to a large extent.

# graphical  
# barplot  
ggplot(data,aes(x=Age,fill=Department\_Name)) +  
 geom\_bar(position="dodge") + facet\_wrap(~Department\_Name) + scale\_fill\_viridis\_d() +  
 ylab("Number of reviews")



# density plot  
ggplot(data,aes(x=Age,fill=Department\_Name)) +  
 geom\_density(position="dodge") + facet\_wrap(~Department\_Name) + scale\_fill\_viridis\_d() +  
 ylab("Density of reviews")



##### To numberically document the distributions, we bind the statistics into one table as follows. We can tell that even though we claim that the distributions are very similar according to the graphs, they have various means and medians. For instance, Tops’ mean age surpasses that of Intimate by almost 3 years old. Nonetheless, these nuances are negligible and we can state that the distribution of age of reviews does not quite vary across product departments.

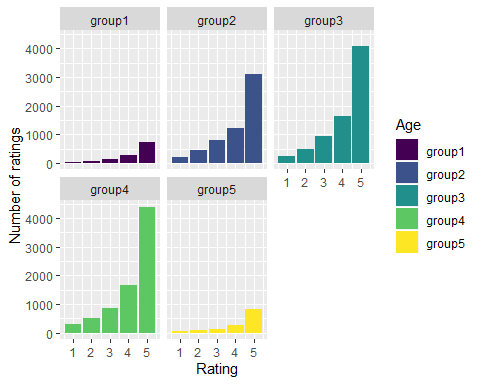
# numerical  
departments = c("Bottoms","Dresses","Intimate","Jackets","Tops","Trend")  
  
output <- NULL  
  
for(i in departments){  
 num\_summary <- data[c("Age","Department\_Name")] %>%  
 filter(Department\_Name==i) %>% summarise(Department=i,  
 Avg = mean(Age),  
 Med = median(Age),  
 '25%ile' = quantile(Age,0.25),  
 '75%ile' = quantile(Age,0.75),  
 StD = sd(Age),  
 IQR = IQR(Age)  
 )  
 lil\_tb <- tbl\_df(num\_summary)  
 output <- bind\_rows(lil\_tb,output)  
}  
  
print(output)

## # A tibble: 6 x 7  
## Department Avg Med `25%ile` `75%ile` StD IQR  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Trend 44.1 43 36 53 12.3 17   
## 2 Tops 44.1 42 35 53 12.5 18   
## 3 Jackets 44.0 42 34 53 13.0 19   
## 4 Intimate 41.3 39 32.5 49 12.3 16.5  
## 5 Dresses 42.1 40 33 50 12.0 17   
## 6 Bottoms 43.1 41 35 51 11.8 16

### Q2

##### In this procedure, for analytical purposes, we divide respondent age into five demographic categories: 25 and under, 26 - 35, 36-45, 46-64, and 65 and over. We use “group1”, “group2”, “group3”, “group4” and “group5” to represent them, respectively. We graph these groups with barplots w.r.t product ratings. We note that although they are different vertically under the same scale, the distribution patterns are generally identical. Although group3’s and group4’s average ratings are smaller than those of some other groups, the people in these age groups contribute the most to the number of (purchases and) reviews, thus people of age 36-45 and 46-64 are the most enthusiastic towards this company’s products.

# modified data: data\_copy  
data\_copy <- data  
data\_copy$Age[data\_copy$Age>=65] <- "group5"  
data\_copy$Age[data\_copy$Age>=46 & data\_copy$Age<=64] <- "group4"  
data\_copy$Age[data\_copy$Age>=36 & data\_copy$Age<=45] <- "group3"  
data\_copy$Age[data\_copy$Age>=26 & data\_copy$Age<=35] <- "group2"  
data\_copy$Age[data\_copy$Age<=25] <- "group1"  
  
# graphical  
ggplot(data\_copy,aes(x=Rating,fill=Age)) +  
 geom\_bar(position="dodge") + facet\_wrap(~Age) + scale\_fill\_viridis\_d() +  
 ylab("Number of ratings")



# numerical  
groups = c("group1","group2","group3","group4","group5")  
  
output <- NULL  
  
for(i in groups){  
 num\_summary <- data\_copy[c("Age","Rating")] %>%  
 filter(Age==i) %>% summarise(Department=i,Avg = mean(Rating),  
 Med = median(Rating),  
 '25%ile' = quantile(Rating,0.25),  
 '75%ile' = quantile(Rating,0.75),  
 StD = sd(Rating),  
 IQR = IQR(Rating)  
 )  
 lil\_tb <- tbl\_df(num\_summary)  
 output <- bind\_rows(lil\_tb,output)  
}  
  
print(output)

## # A tibble: 5 x 7  
## Department Avg Med `25%ile` `75%ile` StD IQR  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 group5 4.26 5 4 5 1.11 1  
## 2 group4 4.22 5 4 5 1.11 1  
## 3 group3 4.19 5 4 5 1.10 1  
## 4 group2 4.14 5 3 5 1.14 2  
## 5 group1 4.29 5 4 5 1.04 1

# ——————————————————————

# Task 3

### We present three lists of sorted products according to their popularity calculated based on three different criteria. In this case, the third list is the most appropriate list to consider

#### Table 1 (highest average ratings)

# first table: product ID’s with the highest average ratings  
table1 <- data %>% add\_count(Clothing\_ID) %>%  
 group\_by(Clothing\_ID) %>% mutate(Avg\_rate = mean(Rating)) %>% mutate(prop = sum(Recommended)/n) %>%  
 select(c("Clothing\_ID","n","Avg\_rate","prop","Department\_Name")) %>% unique()  
table1 <- table1[order(-table1$Avg\_rate,table1$n),]  
table1[1:10,]

## # A tibble: 10 x 5  
## # Groups: Clothing\_ID [10]  
## Clothing\_ID n Avg\_rate prop Department\_Name  
## <dbl> <int> <dbl> <dbl> <chr>   
## 1 4 1 5 1 Tops   
## 2 1196 1 5 1 Dresses   
## 3 329 1 5 1 Intimate   
## 4 596 1 5 1 Trend   
## 5 1182 1 5 1 Tops   
## 6 565 1 5 1 Trend   
## 7 580 1 5 1 Intimate   
## 8 234 1 5 1 Intimate   
## 9 204 1 5 1 Intimate   
## 10 548 1 5 1 Trend

#### Table 2 (highest proportion of positive recommendations)

# second table: product ID’s with the highest proportion of positive recommendations  
table2 <- table1  
table2 <- table1[order(-table1$prop,table1$n),]  
table2[1:10,]

## # A tibble: 10 x 5  
## # Groups: Clothing\_ID [10]  
## Clothing\_ID n Avg\_rate prop Department\_Name  
## <dbl> <int> <dbl> <dbl> <chr>   
## 1 4 1 5 1 Tops   
## 2 1196 1 5 1 Dresses   
## 3 329 1 5 1 Intimate   
## 4 596 1 5 1 Trend   
## 5 1182 1 5 1 Tops   
## 6 565 1 5 1 Trend   
## 7 580 1 5 1 Intimate   
## 8 234 1 5 1 Intimate   
## 9 204 1 5 1 Intimate   
## 10 548 1 5 1 Trend

#### Table 3 (highest Wilson lower confidence limits)

# third table: product ID’s with the highest Wilson lower confidence limits for positive recommendations as described above  
  
# helper function  
helper <- function(x2,x4){  
 # x1 is the Clothing\_ID  
 # x2 is the n  
 # x3 is the Avg\_rate  
 # x4 is the prop  
 # x5 is the Department\_Name  
 a = 1.96 \* 1.96 / (2 \* x2)  
 b = x4 \* (1 - x4) / x2  
 c = a / (2 \* x2)  
 WLCL = (x4 + a - 1.96 \* sqrt(b + c)) / (1 + 2 \* a)  
 return(WLCL)  
}  
  
table3 <- table1 %>% group\_by(Clothing\_ID) %>% mutate(WLCL=helper(n,prop))  
table3 <- table3[order(-table3$WLCL),]  
table3[1:10,]

## # A tibble: 10 x 6  
## # Groups: Clothing\_ID [10]  
## Clothing\_ID n Avg\_rate prop Department\_Name WLCL  
## <dbl> <int> <dbl> <dbl> <chr> <dbl>  
## 1 1123 30 4.7 1 Jackets 0.886  
## 2 834 150 4.54 0.933 Tops 0.882  
## 3 1025 125 4.46 0.936 Bottoms 0.879  
## 4 1008 186 4.46 0.914 Bottoms 0.865  
## 5 984 175 4.46 0.914 Jackets 0.863  
## 6 839 48 4.56 0.958 Tops 0.860  
## 7 1024 35 4.66 0.971 Bottoms 0.855  
## 8 1033 220 4.43 0.895 Bottoms 0.848  
## 9 872 545 4.38 0.877 Tops 0.847  
## 10 1026 21 4.81 1 Bottoms 0.845