Final Project in R

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.

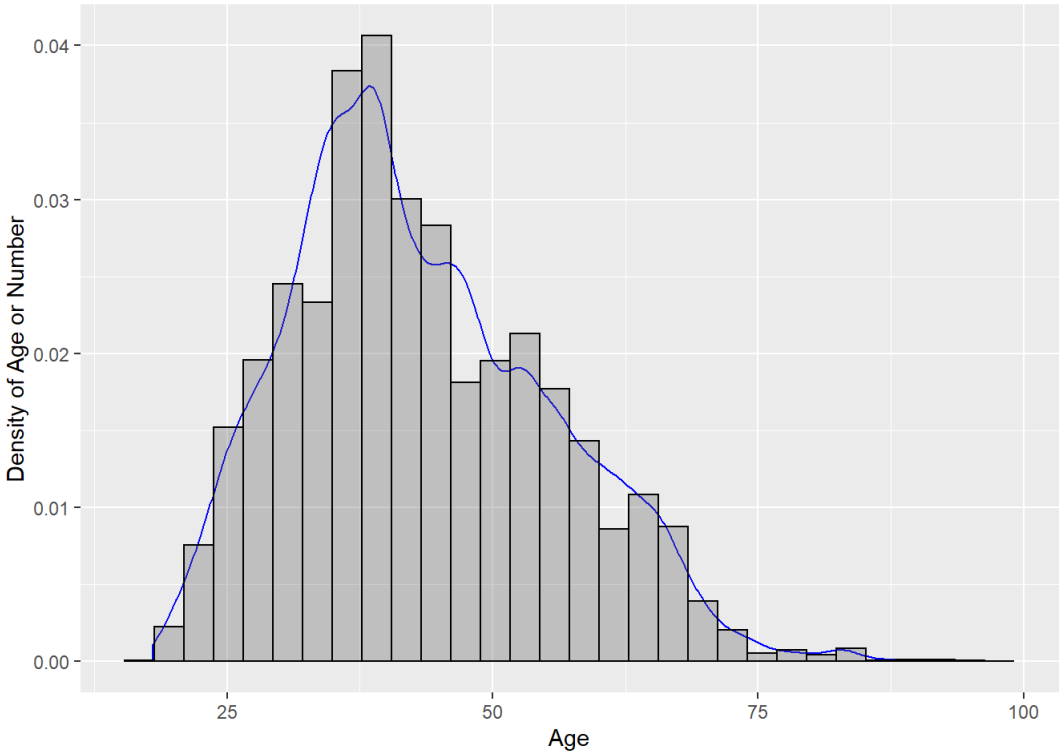
```
# Load dependencies
library(ggplot2)
library(dplyr)
library(tidyverse)
library(forcats)
library(readr)
library(knitr)

# Read data and select fields
Womens_Clothing_Reviews <- read_csv("D:/PERSONAL/McGill/McGill/McGill Current/MATH 208/assignments/DataAnalysis_R/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-histogram 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(col="blue") + geom_histogram(aes(y=..density..),bins=30,col="black",alpha=0.3) + xl
ab("Age") + ylab("Density of Age or Number")
```



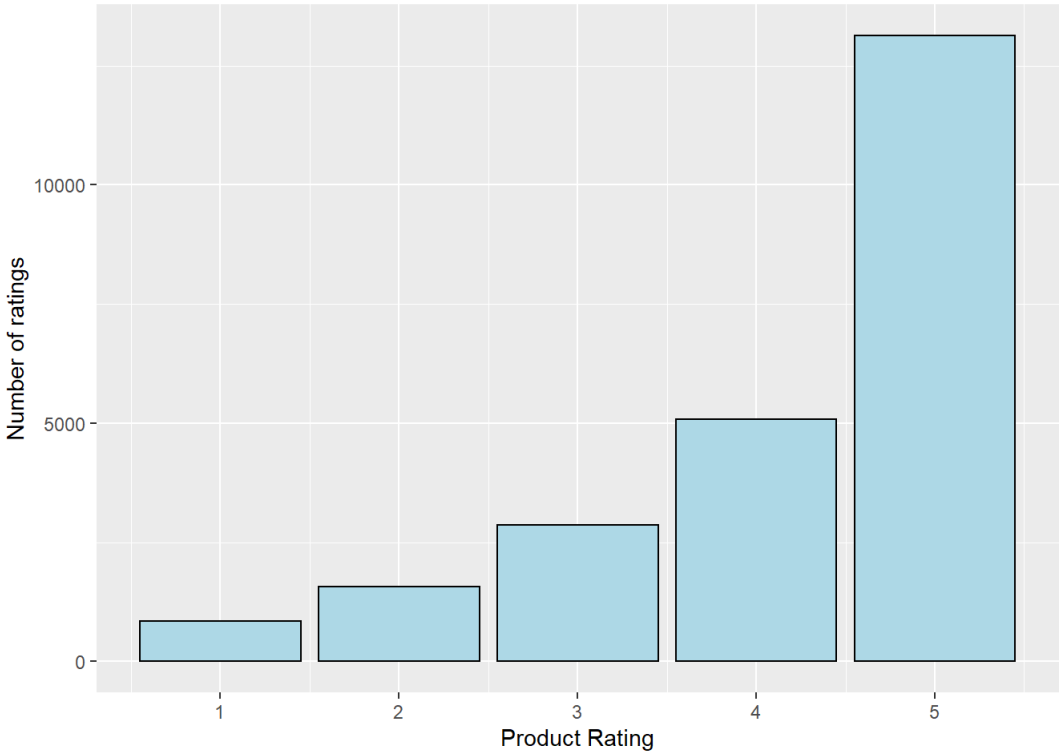

```
# Summary table
data %>% summarise(Avg = mean(Age),
                   Med = median(Age),
                   '25%ile' = quantile(Age,0.25),
                   '75%ile' = quantile(Age,0.75),
                   StD = sd(Age),
                   IQR = IQR(Age)
                   ) %>% kable()
```

Avg	Med	25%ile	75%ile	StD	IQR
43.19854	41	34	52	12.27954	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 drops accordingly.

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




```
# Summary table 1
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)
                   ) %>% kable()
```

Avg	Med	25%ile	75%ile	StD	IQR
4.196032	5	4	5	12.27954	18

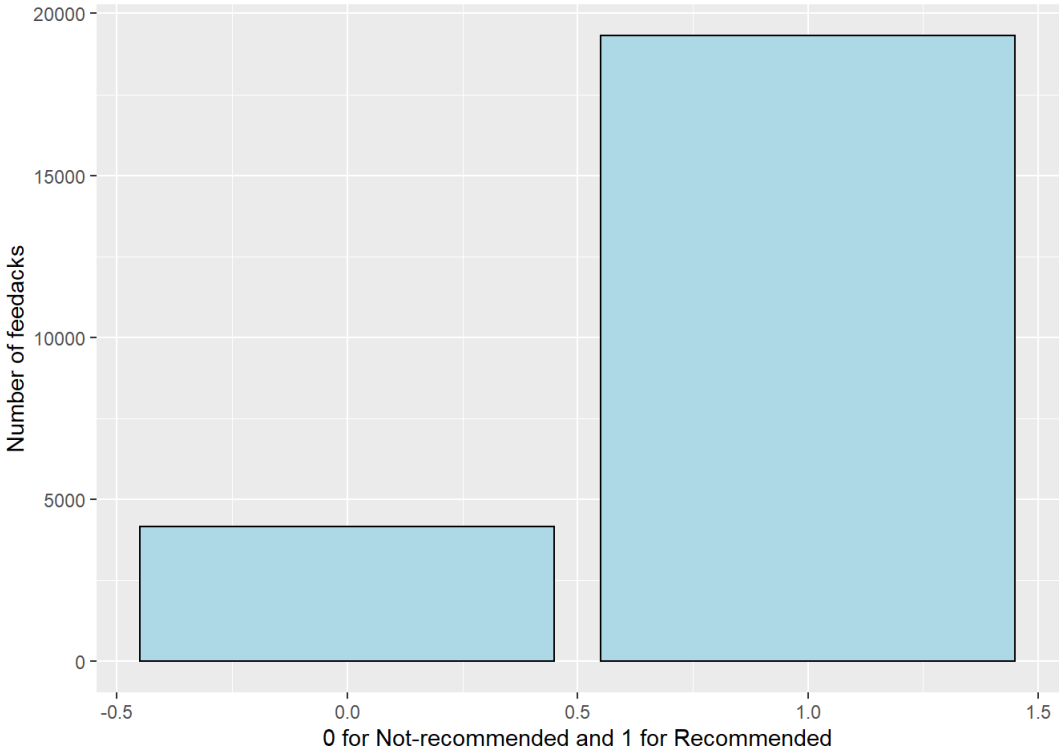
```
# Summary table 2
data %>% group_by(Rating) %>% summarise(count=n()) %>%
mutate(prop=count/sum(count)) %>% arrange(desc(prop)) %>% kable()
```

Rating	count	prop
5	13131	0.5590990
4	5077	0.2161713
3	2871	0.1222430
2	1565	0.0666354
1	842	0.0358511

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")
```



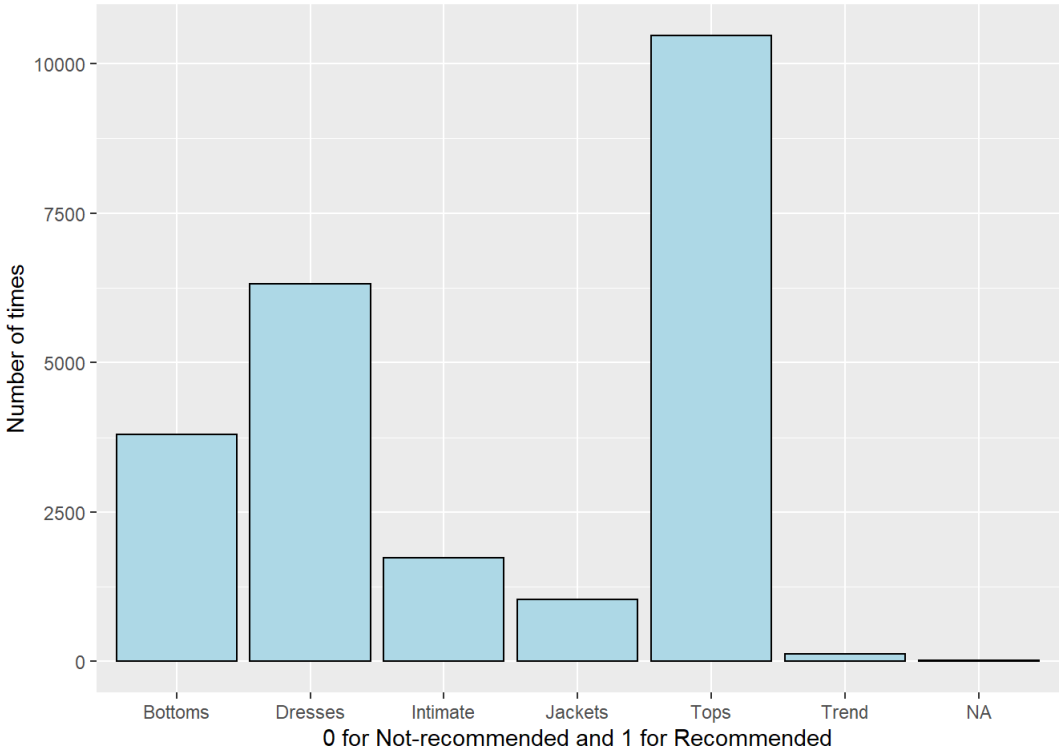

```
# helper
counter <- function(x){
  a <- table(data$Recommended)
  b <- a[names(a)==x]
  return(b[1])
}

# 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)
                   ) %>% kable()
```

Avg	Med	25%ile	75%ile	StD	IQR	NotRecommended	Recommended
0.8223623	1	1	1	0.3822156	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 %>% group_by(Department_Name) %>% summarise(count=n()) %>%
mutate(prop=count/sum(count)) %>% arrange(desc(prop)) %>% kable()
```

Department_Name	count	prop
Tops	10468	0.4457123
Dresses	6319	0.2690539
Bottoms	3799	0.1617559
Intimate	1735	0.0738738
Jackets	1032	0.0439411
Trend	119	0.0050668
NA	14	0.0005961

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),]
dim(data)
```

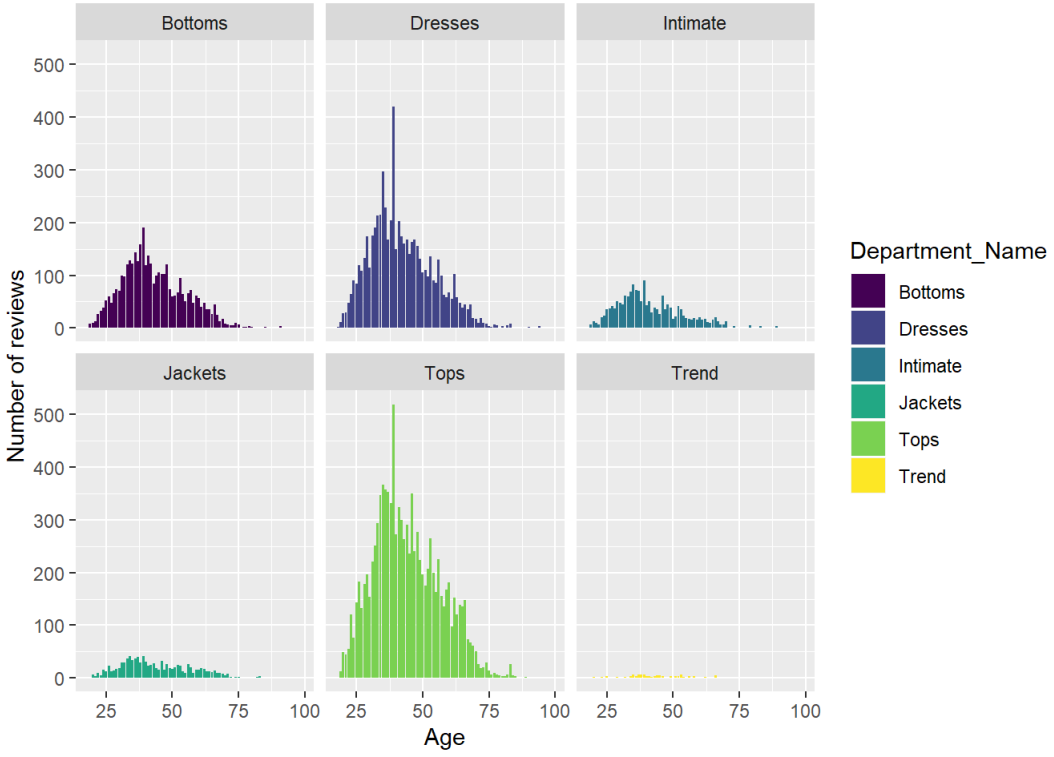
```
## [1] 23472     6
```

Task 2

Q1

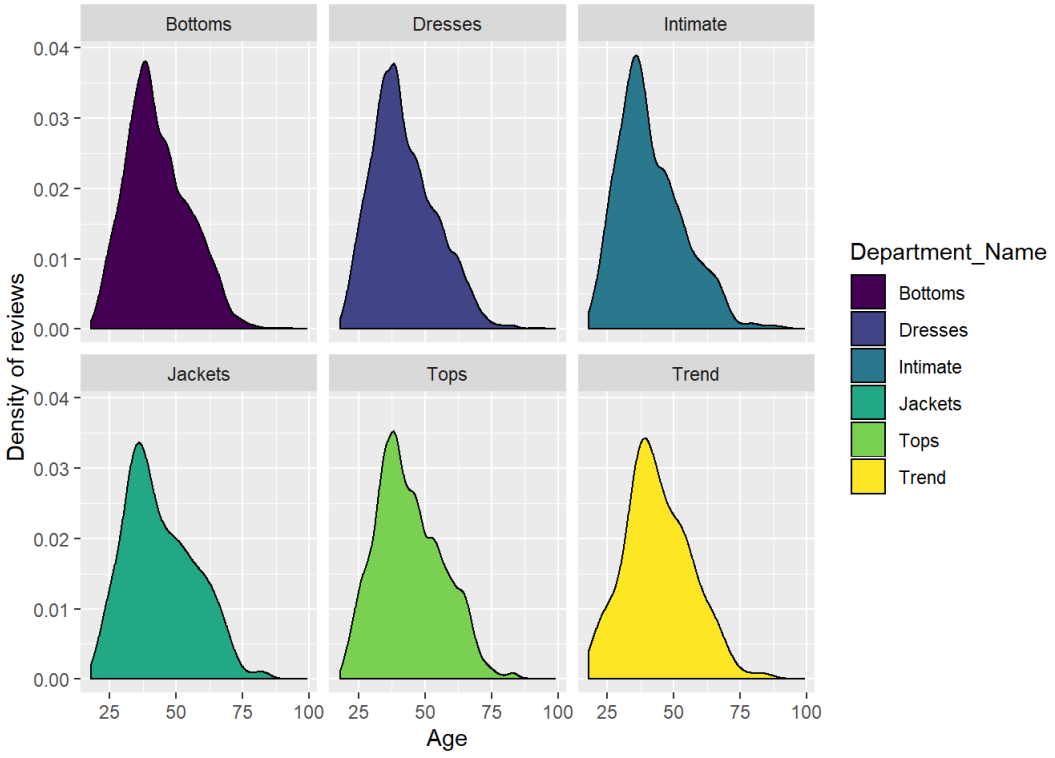
If we separate the six departments apart w.r.t. to Age as in the following diagrams, we can tell that the six departments' products are not equally popular under one scale in the barplot: Tops and Dresses received the largest number of reviews, whereas Trend the smallest, which is almost unnoticeable.

```
# 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")
```



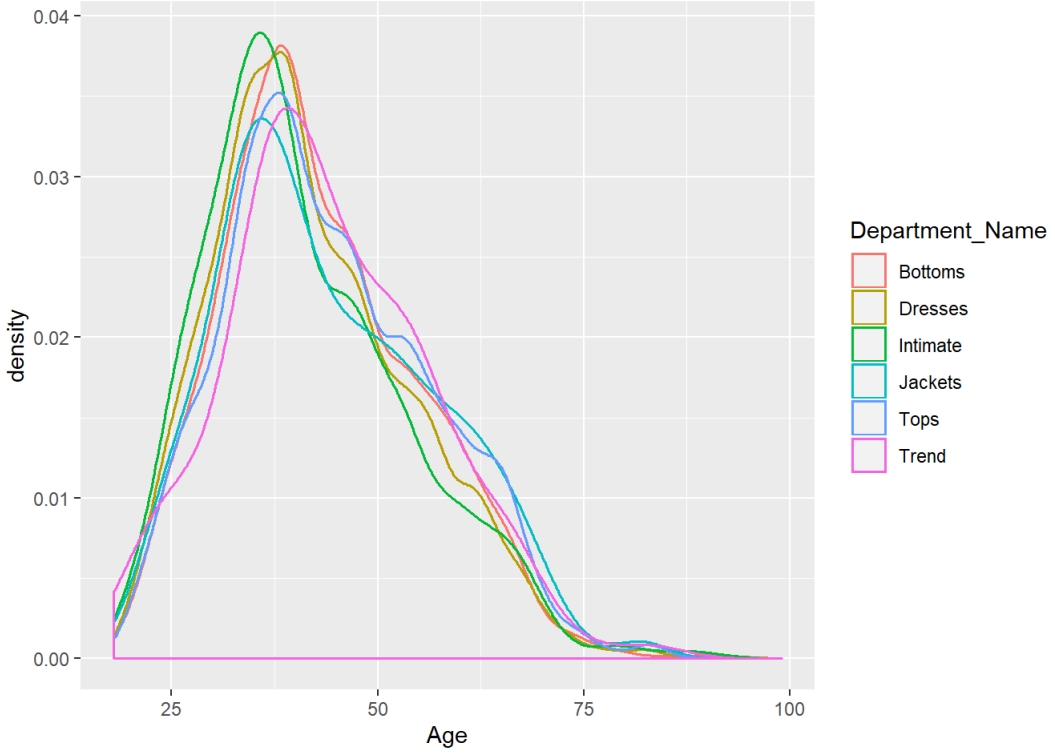
However, if we graph the same data in a set of density plots, we discover that the distributions of the six departments' product reviews are similar to a large extent.

```
# 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")
```



Our proposition can be further confirmed if we graph they six onto one plot as follows.

```
ggplot(data,aes(x=Age,col=Department_Name)) +
geom_density(size=0.6)
```



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)
}

kable(output)
```

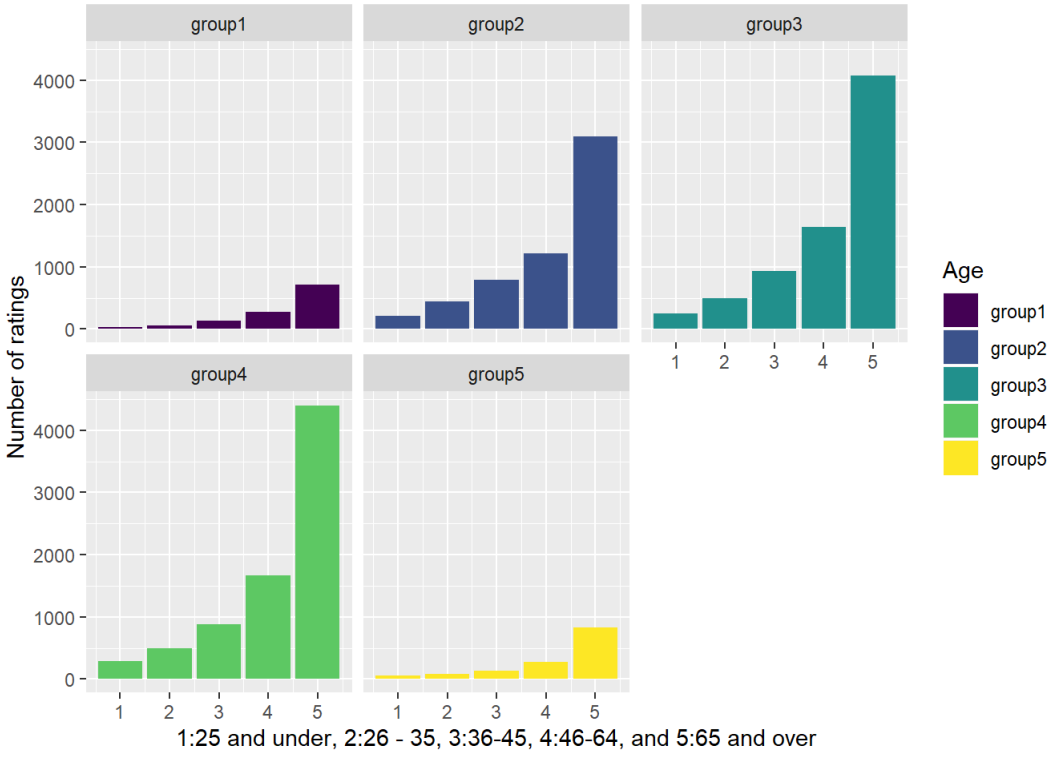
Department	Avg	Med	25%ile	75%ile	StD	IQR
Trend	44.05882	43	36.0	53	12.27944	17.0
Tops	44.12591	42	35.0	53	12.46187	18.0
Jackets	43.96415	42	34.0	53	13.04769	19.0
Intimate	41.29568	39	32.5	49	12.34064	16.5
Dresses	42.11489	40	33.0	50	11.96692	17.0
Bottoms	43.09318	41	35.0	51	11.79964	16.0

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") + xlab("1:25 and under, 2:26 - 35, 3:36-45, 4:46-64, and 5:65 and over")
```




```
# 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(output,lil_tb)
}

kable(output)
```

Department	Avg	Med	25%ile	75%ile	StD	IQR
group1	4.291564	5	4	5	1.037100	1
group2	4.136150	5	3	5	1.138756	2
group3	4.190495	5	4	5	1.100438	1
group4	4.217262	5	4	5	1.106718	1
group5	4.263348	5	4	5	1.111436	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. This is because the algorithm behind the

calculation of the list balances both the number of reviews with the recommended proportion instead of relying on solely the average review or the proportion recommended. Specifically, the products chosen by the first two algorithms all have perfect average ratings or perfect proportion scores of recommendation, but, more importantly, are reviewd for only once. On the contrary, the top ten products selected by the third list are generally products with a considerable number of reviews (i.e., purchases) and good scores (not necessarily perfect) of average ratings and proportion. Obviously, the result of the third list is more inclusive and reasonable, and this algorithm should be given the first priority.

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),]
kable(table1[1:10,])
```

Clothing_ID	n	Avg_rate	prop	Department_Name
4	1	5	1	Tops
1196	1	5	1	Dresses
329	1	5	1	Intimate
596	1	5	1	Trend
1182	1	5	1	Tops
565	1	5	1	Trend
580	1	5	1	Intimate
234	1	5	1	Intimate
204	1	5	1	Intimate
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),]
kable(table2[1:10,])
```

Clothing_ID	n	Avg_rate	prop	Department_Name
4	1	5	1	Tops
1196	1	5	1	Dresses
329	1	5	1	Intimate
596	1	5	1	Trend
1182	1	5	1	Tops
565	1	5	1	Trend
580	1	5	1	Intimate
234	1	5	1	Intimate
204	1	5	1	Intimate
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 abov
e

# 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),]
kable(table3[1:10,])
```

Clothing_ID	n	Avg_rate	prop	Department_Name	WLCL
1123	30	4.700000	1.0000000	Jackets	0.8864829
834	150	4.540000	0.9333333	Tops	0.8816365
1025	125	4.464000	0.9360000	Bottoms	0.8787832
1008	186	4.462366	0.9139785	Bottoms	0.8648440
984	175	4.462857	0.9142857	Jackets	0.8634038
839	48	4.562500	0.9583333	Tops	0.8602409
1024	35	4.657143	0.9714286	Bottoms	0.8546659
1033	220	4.427273	0.8954545	Bottoms	0.8480142
872	545	4.383486	0.8770642	Tops	0.8468267
1026	21	4.809524	1.0000000	Bottoms	0.8453562