MATH 208 Final Project

Yan Miao

2019-11-20

Introduction

Understanding customer needs is of paramount importance in marketing strategies today. Not only will it give companies an insight as to how customers perceive their products and/or services, but it will also give them an idea on how to improve their offers. This project attempts to perform data analysis and understand the correlation of different variables in customer reviews on a women clothing e-commerce.These goals are achieved by completing three sub-tasks. We first explore the each single variable by plotting out their distributions and presenting the summary tables. Then, we dig deeper into the data by analyzing the associations between two pairs of variables: **Age** & **Department_Name** and **Age** & **Rating**. This will provide valuable information for future marketing strategies targetting at people of different ages. Finally, we aimed to find out the best way to select out the most popular products based on the previous ratings so that we could better understand what customers need.

Preparation

Load libraries:

```
library(tidyverse)
library(gridExtra)
library(kableExtra)
library(lemon) # plotting table
library(stringr) # matching strings
knit_print.data.frame <- lemon_print
```

```
WCR<-read_csv("Womens_Clothing_Reviews.csv")
head(WCR)
```

```
## # A tibble: 6 x 11
##   Review_ID Clothing_ID   Age Title Review_Text Rating Recommended
##       <dbl>       <dbl> <dbl> <chr> <chr>        <dbl>       <dbl>
## 1         0         767    33 <NA>  Absolutely~      4           1
## 2         1        1080    34 <NA>  "Love this~      5           1
## 3         2        1077    60 Some~ I had such~      3           0
## 4         3        1049    50 My f~ I love, lo~      5           1
## 5         4         847    47 Flat~ This shirt~      5           1
## 6         5        1080    49 Not ~ I love tra~      2           0
## # ... with 4 more variables: Positive_Feedback_Count <dbl>,
## #   Division_Name <chr>, Department_Name <chr>, Class_Name <chr>
```

Selecting out the columns we need:

```
WCR<-WCR %>% select("Review_ID","Clothing_ID","Age","Rating","Recommended","Department_Name")
head(WCR)
```

Review_ID	Clothing_ID	Age	Rating	Recommended	Department_Name
0	767	33	4	1	Intimate
1	1080	34	5	1	Dresses
2	1077	60	3	0	Dresses
3	1049	50	5	1	Bottoms
4	847	47	5	1	Tops
5	1080	49	2	0	Dresses

Task 1: Exploratory single variable analyses

Check if there exists any missing values

```
sum(is.na(WCR))
```

```
## [1] 14
```

Remove rows with missing values

```
WCR<-na.omit(WCR)
sum(is.na(WCR))
```

```
## [1] 0
```

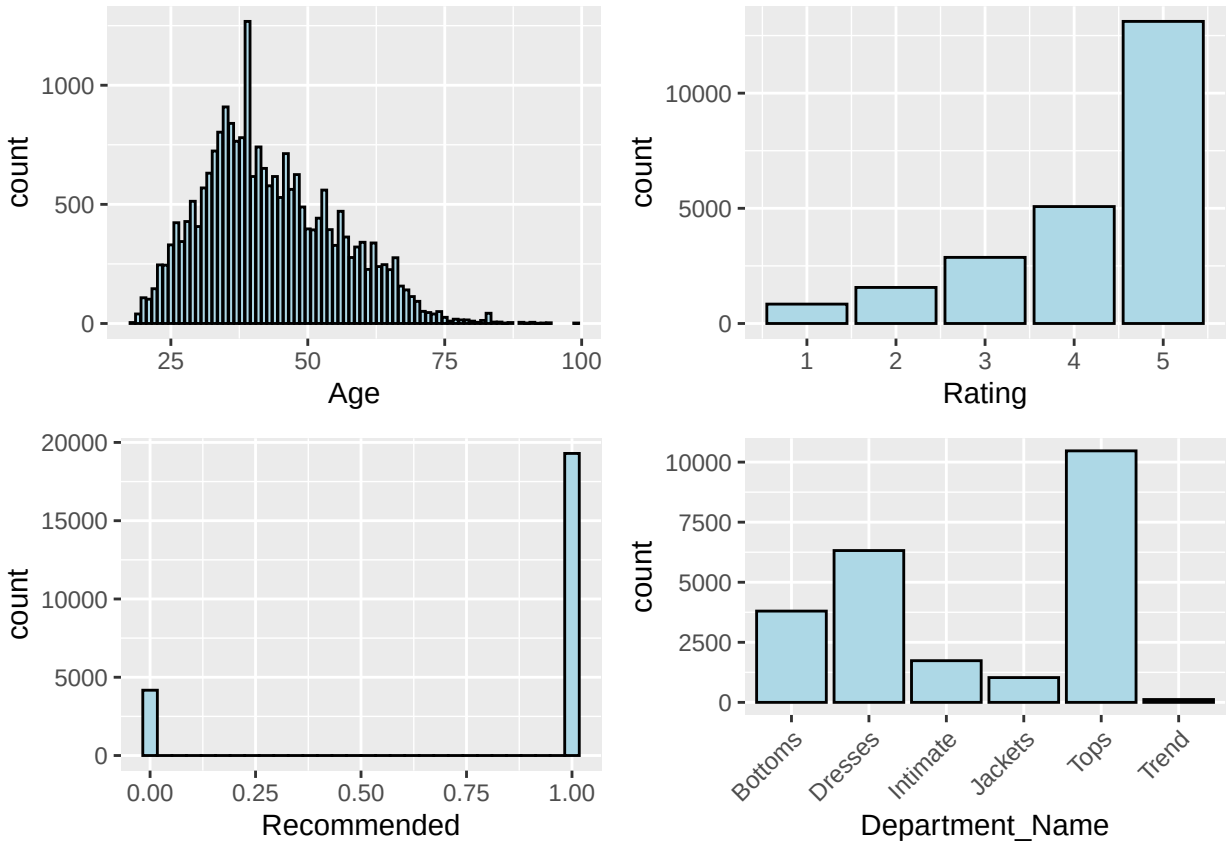
Data exploration:

```
summary(WCR)
```

Review_ID	Clothing_ID	Age	Rating	Recommended	Department_Name
Min. : 0	Min. : 0.0	Min. :18.0	Min. :1.000	Min. :0.0000	Length:23472
1st Qu.: 5868	1st Qu.: 861.0	1st Qu.:34.0	1st Qu.:4.000	1st Qu.:1.0000	Class :character
Median :11736	Median : 936.0	Median :41.0	Median :5.000	Median :1.0000	Mode :character
Mean :11739	Mean : 918.5	Mean :43.2	Mean :4.196	Mean :0.8223	
3rd Qu.:17610	3rd Qu.:1078.0	3rd Qu.:52.0	3rd Qu.:5.000	3rd Qu.:1.0000	
Max. :23485	Max. :1205.0	Max. :99.0	Max. :5.000	Max. :1.0000	

Plot distributions:

```
p1<-ggplot(WCR, aes(x = Age)) + geom_bar(col = 'black',fill="lightblue")
p2<-ggplot(WCR, aes(x = Rating)) + geom_bar(col = 'black',fill="lightblue")
p3<-ggplot(WCR, aes(x = Recommended)) + geom_histogram(col = 'black',fill="lightblue")
p4<-ggplot(WCR, aes(x = Department_Name)) + geom_bar(col = 'black',fill="lightblue") +
  theme(axis.text.x = element_text(angle = 45,hjust = 1))
grid.arrange(grobs=list(p1,p2,p3,p4))
```



- **"Age" description**

The "Age" variable is an integer variable. The majority of it distributes in between 25 and 50, with mean being 43.2, median being 41.0. The minimum value is 18.0 and the maximum value is 99.0. It resembles a positively skewed normal distribution. The most common three ages are (the first row represents ages and the second represents number of reviews):

```
sort(table(WCR$Age),decreasing=TRUE)[1:3]
```

```
##
##   39   35   36
## 1268  909  840
```

- **"Rating" description**

The "Rating" variable is an integer variable. It can be chosen from the set: {1,2,3,4,5}. When the value becomes larger, the number of reviews increases accordingly. The mean value for "Rating" is 4.196, which is quite high. The number of reviews under each value of "Rating" is listed below:

```
sort(table(WCR$Rating),decreasing=TRUE)
```

```
##
##     5     4     3     2     1
## 13117  5077  2871  1565   842
```

- **"Recommended" description**

The "Recommeneded" variable is binary with majority being 1. This means that most of the customers are satisfied with the products they purchased, and thus, most of the reviews are positive. The ratio of positive to negative reviews is 19300:4172.

- **"Department_Name" description**

Different from the other variables, the "Department_Name" variable is a categorical variable. The values are chosen from the set: {"Intimate","Dresses","Bottoms","Tops","Jackets","Trend"}. The majority of the reviews are about the products from the "Tops" department while the minority of them are about the products from the "Trend" department.

Task 2: Exploring associations

Question 1

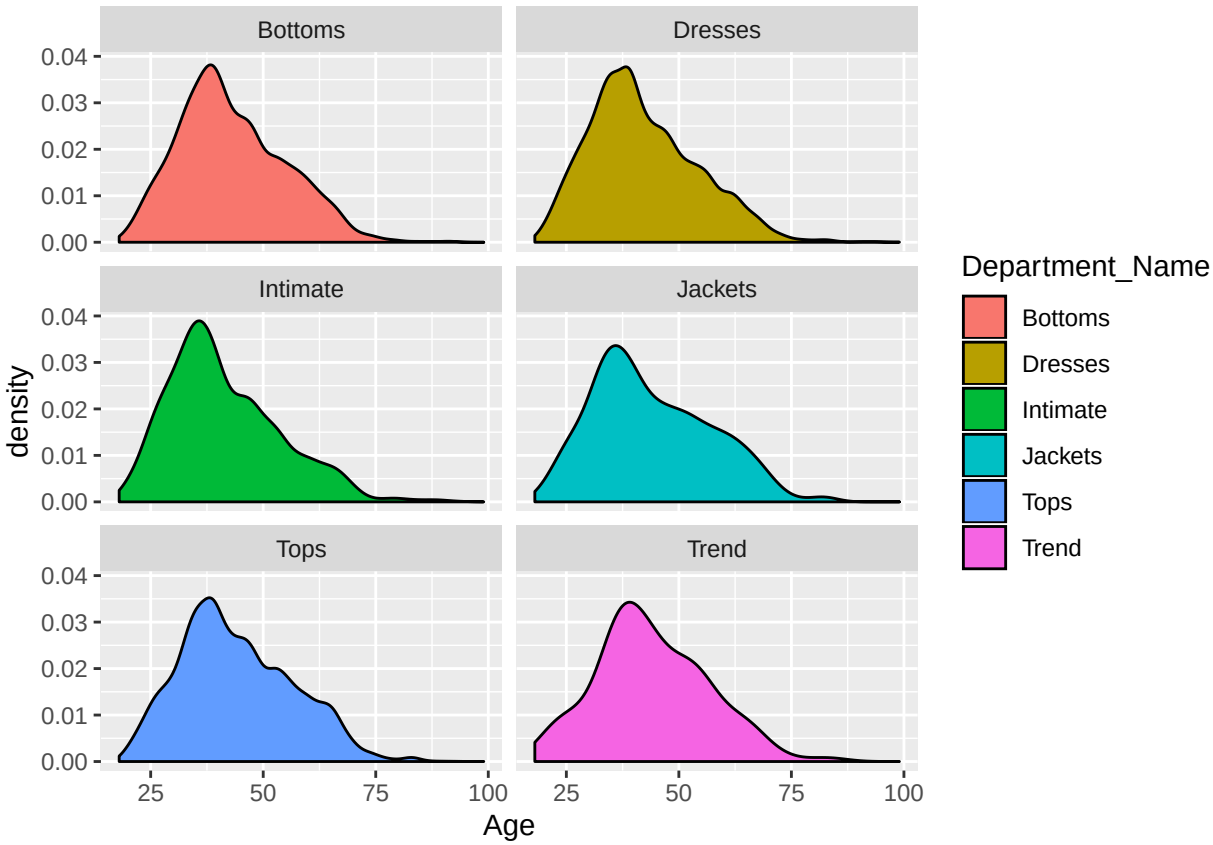
Summary table:

```
WCR %>% group_by(Department_Name) %>% summarise(Avg = mean(Age),
                                                Med = median(Age),
                                                Std = sd(Age),
                                                count = n())
```

Department_Name	Avg	Med	Std	count
Bottoms	43.09318	41	11.79964	3799
Dresses	42.11489	40	11.96692	6319
Intimate	41.29568	39	12.34064	1735
Jackets	43.96415	42	13.04769	1032
Tops	44.12591	42	12.46187	10468
Trend	44.05882	43	12.27944	119

Plot distribution

```
ggplot(WCR, aes(x = Age, group = Department_Name, fill = Department_Name)) +
  geom_density() +
facet_wrap(~Department_Name,nrow = 3, ncol = 2)
```



From the plots and the table above, we can see that although the total number of reviewers varies across different departments (with the most reviews for "Tops" department and the fewest for "Trend" department), the majority of the reviewers are of 25-50 years of age. Also, the average ages for each of these 6 groups are within 41 to 45, which are quite close to each other. And the same applies to the median and the standard deviation.

Question 2

```
age_frame<-WCR %>% mutate(age_cat = ifelse(Age <= 25, '25 and under',
                                           ifelse(Age <= 35, '26-35',
                                                  ifelse(Age <= 45, '36-45',
                                                         ifelse(Age <= 64, '46-64',
                                                                '65 and over')))))

age_frame<-age_frame %>% mutate(age_cat = factor(age_cat),
                                Rating = factor(Rating, levels = c('1','2','3','4','5')))
```

```
num_ppl<-age_frame %>% group_by(age_cat) %>% count(Rating) %>% .["n"]
ppl_frame<-matrix(unlist(num_ppl),ncol=5,byrow=TRUE)
colnames(ppl_frame) <- c("1","2","3","4","5")
rownames(ppl_frame) <- c("25 and under","26-35","36-45","46-64","65 and over")
ppl_frame %>% kable(caption = "Number of Reviews under different ages and ratings")
```

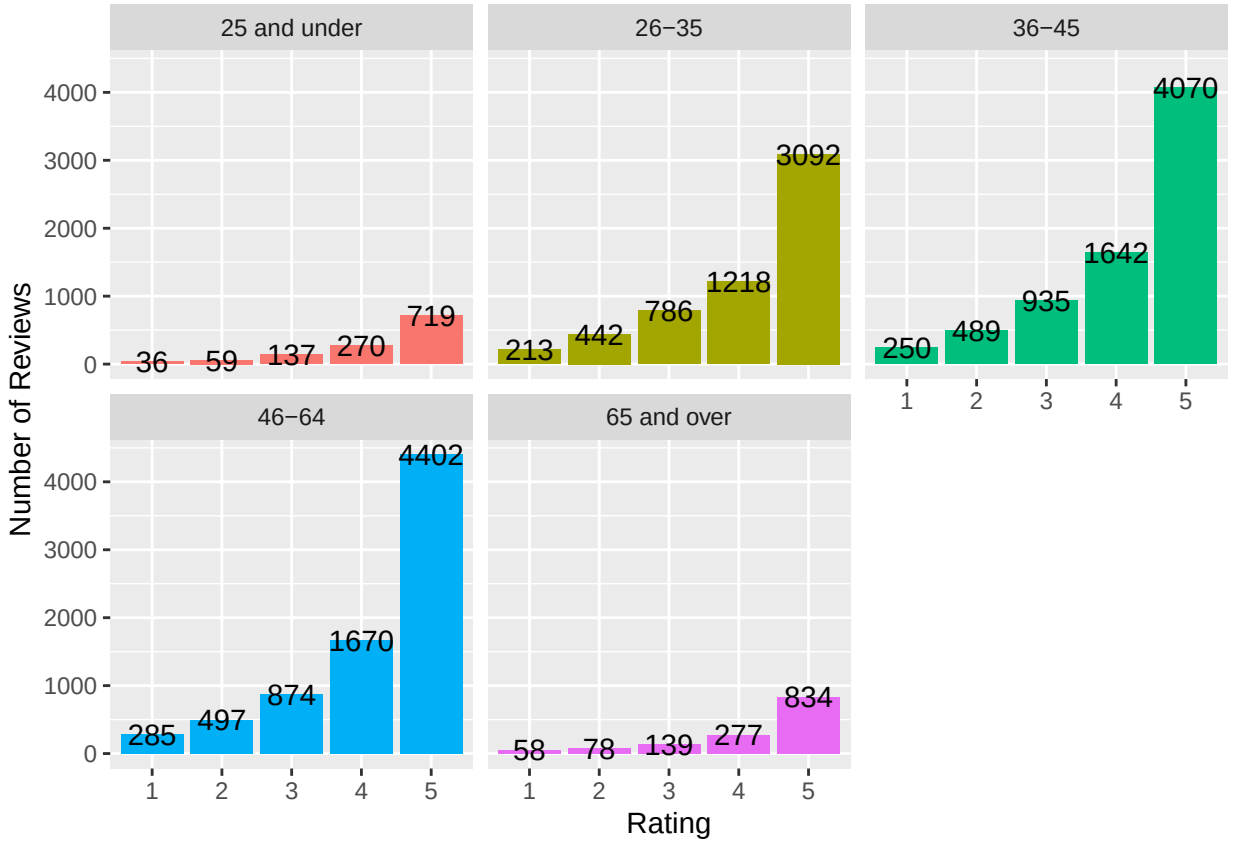
See Table 1.

```
age_frame %>% group_by(age_cat) %>% count(Rating) %>%
  ggplot(aes(x = Rating,y = n, fill = age_cat)) +
```

Table 1: Number of Reviews under different ages and ratings

	1	2	3	4	5
25 and under	36	59	137	270	719
26-35	213	442	786	1218	3092
36-45	250	489	935	1642	4070
46-64	285	497	874	1670	4402
65 and over	58	78	139	277	834

```
geom_bar(stat='identity', show.legend = FALSE) + facet_wrap(~age_cat) +
xlab('Rating') + ylab('Number of Reviews') + geom_text(aes(label = n))
```



The distributions look similar to each other, with the number of reviews increasing when the rating goes higher. To "measure" the enthusiasm for each group, we calculate the percentage of "5" under each rating group as follows.

$$Enthusiasm(\text{"25 and under"}) = \frac{719}{719+270+137+59+36} = \frac{719}{1221} \approx 58.9\%$$
$$Enthusiasm(\text{"26-35"}) = \frac{3092}{3092+1218+786+442+213} = \frac{3092}{5751} \approx 53.8\%$$
$$Enthusiasm(\text{"36-45"}) = \frac{4070}{4070+1642+935+489+250} = \frac{4070}{7386} \approx 55.1\%$$
$$Enthusiasm(\text{"46-64"}) = \frac{4402}{4402+1670+874+497+285} = \frac{4402}{7728} \approx 57.0\%$$
$$Enthusiasm(\text{"65 and over"}) = \frac{834}{834+277+139+78+58} = \frac{834}{1386} \approx 60.2\%$$

More than half of the reviewers from each group rated 5 for the products, which means products from this company are quite popular in general. Particularlly, the groups "65 and over", "25 and under", and "46-64" have higher top ratings percentage than the others and the group "65 and over" has more than 60 percent of the reviews giving a highest rating. This shows that these groups are the most enthusiastic about the company's products.

Task 3

Part 1

The 10 product ID's with the highest average ratings

```
info1<-WCR %>% group_by(Clothing_ID) %>% summarise(Avg_rating = mean(Rating)) %>%
  arrange(desc(Avg_rating)) %>% head(10)

dept1<-WCR %>% select(Clothing_ID,Department_Name) %>% group_by(Clothing_ID) %>%
  .[!duplicated(.$Clothing_ID),] %>% filter(Clothing_ID %in% pull(info1, Clothing_ID)) %>%
  arrange(Clothing_ID)

proportion1<-WCR %>% group_by(Clothing_ID,Recommended) %>% summarise(count=n()) %>%
  mutate(prop=count/sum(count)) %>% filter(Recommended==1,Clothing_ID %in% pull(info1, Clothing_ID))

num_review1<-WCR %>% group_by(Clothing_ID) %>%
  filter(Clothing_ID %in% pull(info1, Clothing_ID)) %>% count(.) %>% .["n"]

info1 %>% add_column(.,prop=pull(proportion1,prop),
                     num_positive_review=pull(proportion1,count),
                     num_review=unlist(num_review1),
                    Dept=pull(dept1,Department_Name)) %>% arrange(desc(Avg_rating)) %>%
  .[,c("Clothing_ID","Dept","num_positive_review","num_review","prop","Avg_rating")]
```

Clothing_ID	Dept	num_positive_review	num_review	prop	Avg_rating
0	Jackets	1	1	1	5
3	Tops	1	1	1	5
4	Tops	1	1	1	5
5	Tops	1	1	1	5
6	Tops	1	1	1	5
7	Jackets	1	1	1	5
12	Tops	1	1	1	5
14	Intimate	1	1	1	5
16	Dresses	1	1	1	5
17	Dresses	1	1	1	5

But in fact, the number of products with an average rating of 5 is

```
WCR %>% group_by(Clothing_ID) %>% summarise(Avg_rating = mean(Rating)) %>%
  filter(Avg_rating==5) %>% count(.)
```

n
377

Apparently, employing the average rating is not performing well for choosing the popular products since there exists a large number of products that have an average rating to be 5.

Part 2

The 10 product ID's with the highest proportion of positive recommendations

```
info2<-WCR %>% group_by(Clothing_ID,Recommended) %>% summarise(count=n()) %>%
  mutate(prop=count/sum(count)) %>% filter(Recommended==1) %>%
  arrange(desc(prop)) %>% head(10)

dept2<-WCR %>% select(Clothing_ID,Department_Name) %>% group_by(Clothing_ID) %>%
  .[!duplicated(.$Clothing_ID),] %>% filter(Clothing_ID %in% pull(info2, Clothing_ID)) %>%
  arrange(Clothing_ID)

avg_rating_col2<-WCR %>% group_by(Clothing_ID) %>% summarise(Avg_rating = mean(Rating)) %>%
  filter(Clothing_ID %in% pull(info2, Clothing_ID))

num_review2<-WCR %>% group_by(Clothing_ID) %>%
  filter(Clothing_ID %in% pull(info2, Clothing_ID)) %>% count(.) %>% .["n"]

info2 %>% add_column(.,Avg_rating=pull(avg_rating_col2,Avg_rating),num_review=unlist(num_review2),
                     Dept=pull(dept2,Department_Name)) %>% arrange(desc(prop)) %>%
  rename(num_positive_review=count) %>%
  .[,c("Clothing_ID","Dept","num_positive_review","num_review","prop","Avg_rating")]
```

Clothing_ID	Dept	num_positive_review	num_review	prop	Avg_rating
0	Jackets	1	1	1	5
2	Tops	1	1	1	4
3	Tops	1	1	1	5
4	Tops	1	1	1	5
5	Tops	1	1	1	5
6	Tops	1	1	1	5
7	Jackets	1	1	1	5
9	Bottoms	1	1	1	4
10	Intimate	1	1	1	4
12	Tops	1	1	1	5

But we notice that the proportion of positive recommendations is also not a good criterion for filtering out the most popular products since we have 681 products with the proportion of positive recommendations being 1.

```
WCR %>% group_by(Clothing_ID,Recommended) %>% summarise(count=n()) %>%
  mutate(prop=count/sum(count)) %>% filter(Recommended==1,prop==1) %>% dim(.)
```

```
## [1] 681   4
```

Part 3

Define functions to compute WLCL

```
f1<-function(n_i){
  return(1.96^2/(2*n_i))
}
f2<-function(p_i,n_i){
  return(p_i*(1-p_i)/n_i)
}
f3<-function(n_i){
  a_i<-f1(n_i)
  return(a_i/(2*n_i))
}
WLCL<-function(p_i,n_i){
  a_i<-f1(n_i)
  b_i<-f2(p_i,n_i)
  c_i<-f3(n_i)
  return((p_i+a_i-1.96*sqrt(b_i+c_i))/(1+2*a_i))
}
```

```
p_i<-WCR %>% group_by(Clothing_ID,Recommended) %>% summarise(count=n()) %>%
  mutate(prop=count/sum(count)) %>% filter(Recommended==1)

n_i<-WCR %>% group_by(Clothing_ID) %>%
  filter(Clothing_ID %in% pull(p_i, Clothing_ID)) %>% count(.) %>% .["n"]

info3<-p_i %>% add_column(.,WLCL=WLCL(pull(p_i,prop),unlist(n_i)),num_review=unlist(n_i)) %>%
  arrange(desc(WLCL)) %>% head(10) %>% arrange(Clothing_ID)

avg_rating_col3<-WCR %>% group_by(Clothing_ID) %>% summarise(Avg_rating = mean(Rating)) %>%
  filter(Clothing_ID %in% pull(info3, Clothing_ID))

dept3<-WCR %>% select(Clothing_ID,Department_Name) %>% group_by(Clothing_ID) %>%
  .[!duplicated(.$Clothing_ID),] %>% filter(Clothing_ID %in% pull(info3, Clothing_ID)) %>%
  arrange(Clothing_ID)

info3 %>% add_column(.,Avg_rating=pull(avg_rating_col3,Avg_rating),
                     Dept=pull(dept3,Department_Name)) %>% arrange(desc(WLCL)) %>%
  rename(num_positive_review=count) %>%
  .[,c("Clothing_ID","Dept","num_positive_review","num_review","prop","Avg_rating","WLCL")]
```

Clothing_ID	Dept	num_positive_review	num_review	prop	Avg_rating	WLCL
1123	Jackets	30	30	1.0000000	4.700000	0.8864829
834	Tops	140	150	0.9333333	4.540000	0.8816365
1025	Bottoms	117	125	0.9360000	4.464000	0.8787832
1008	Bottoms	170	186	0.9139785	4.462366	0.8648440
984	Jackets	160	175	0.9142857	4.462857	0.8634038
839	Tops	46	48	0.9583333	4.562500	0.8602409
1024	Bottoms	34	35	0.9714286	4.657143	0.8546659
1033	Bottoms	197	220	0.8954545	4.427273	0.8480142
872	Tops	478	545	0.8770642	4.383486	0.8468267
1026	Bottoms	21	21	1.0000000	4.809524	0.8453562

By sorting the WLCL scores decreasingly, we finally obtain the list of the top ten popular products. As we can see from the table above, the products with the highest WLCL do not necessarily need to have the proportion of the positive reviews to be 1, and not an average rating of 5. However, the top popular products selected using WLCL score do have relatively high values on both the previous two criterion. Since this lower

confidence limit is the value for which we are 97.5% confident that the true value of the proportion in the population who positively recommend product lies above this quantity, and clearly, it ranks the popularity of the products perfectly, this leads to the conlusion that the list generated by sorting the WLCL score best represents the products which are the most popular.

Conclusions

To sum up, through the above analysis, the company should be aware of the following things:

- Most of their products (from every department) sell to people of age 36 to 64.
- Products from "Tops" department sell the best and products from "Trend" department sell the worst according to the number of reviews.
- People of age 65 and over have the greatest chance of giving a rate of 5 to their products, which indicates that they are the most enthusiastic group of people for their products.
- Use the WLCL score to determine which products are popular.