MATH 208 Final Project

Tianze Lin 260762008

Contents

Read the Data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
```

```
review = read_csv("Womens_Clothing_Reviews.csv")
```

```
## Parsed with column specification:
## cols(
##   Review_ID = col_double(),
##   Clothing_ID = col_double(),
##   Age = col_double(),
##   Title = col_character(),
##   Review_Text = col_character(),
##   Rating = col_double(),
##   Recommended = col_double(),
##   Positive_Feedback_Count = col_double(),
##   Division_Name = col_character(),
##   Department_Name = col_character(),
##   Class_Name = col_character()
## )
```

```
## We only need to focus on 6 variables.
review = review %>% select(Review_ID,Clothing_ID,Age,Rating,Recommended,Department_Name)
head(review)
```

```
## # A tibble: 6 x 6
##   Review_ID Clothing_ID   Age Rating Recommended Department_Name
##       <dbl>       <dbl> <dbl>  <dbl>       <dbl> <chr>
## 1         0         767    33      4           1 Intimate
## 2         1        1080    34      5           1 Dresses
```

```
## 3          2         1077     60       3           0 Dresses
## 4          3         1049     50       5           1 Bottoms
## 5          4          847     47       5           1 Tops
## 6          5         1080     49       2           0 Dresses
```

```
dim(review)
```

```
## [1] 23486     6
```

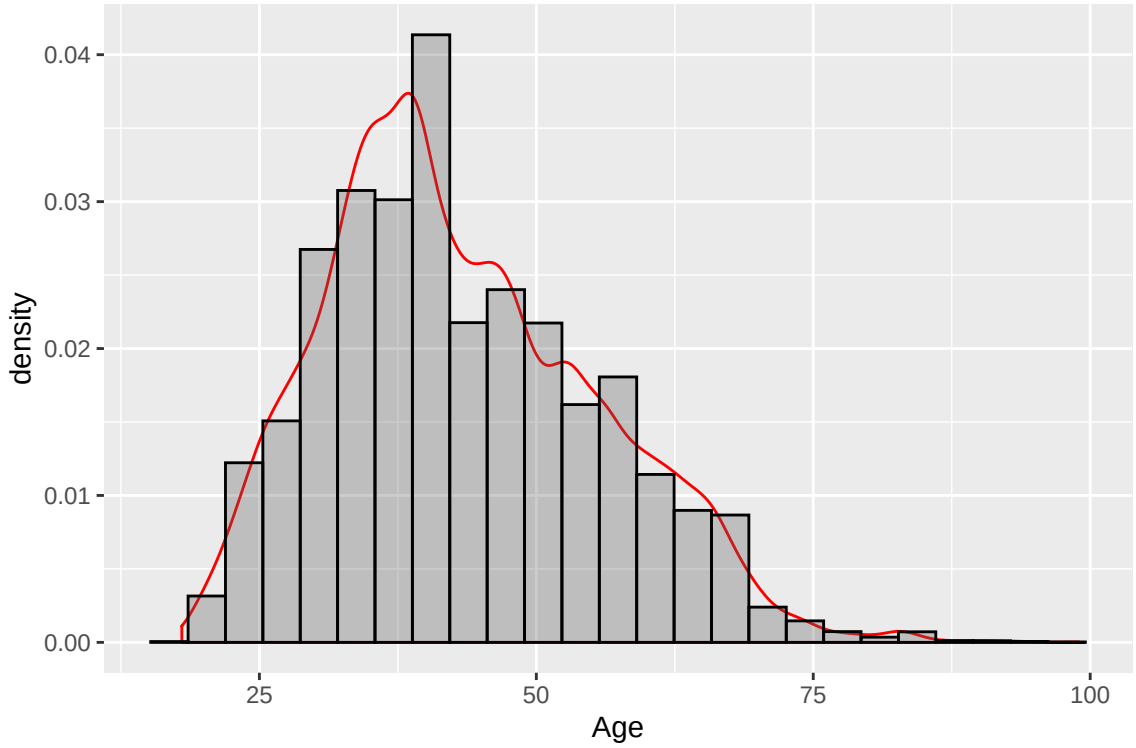
Task 1: Exploratory single variable analyses

In this task, we need to provide some exploratory data analyses and describe the distributions of age, product rating, recommendations, and article departments amongst the respondents. Let's start with Age first.

Age

The Age of the reviewer can be viewed as a continuous random variable. We could use histograms or density plots to estimate its distribution.

```
ggplot(review, aes(x=Age)) + geom_density(col="red")+
  geom_histogram(aes(y=..density..),bins=25,col="black",alpha=0.3)
```




```
review %>% select(Age) %>%
summarise_all(list(Minimum=min,Maximum=max,Average=mean,
                   Median=median,Standard_Deviation=sd)) %>%
  kable()
```

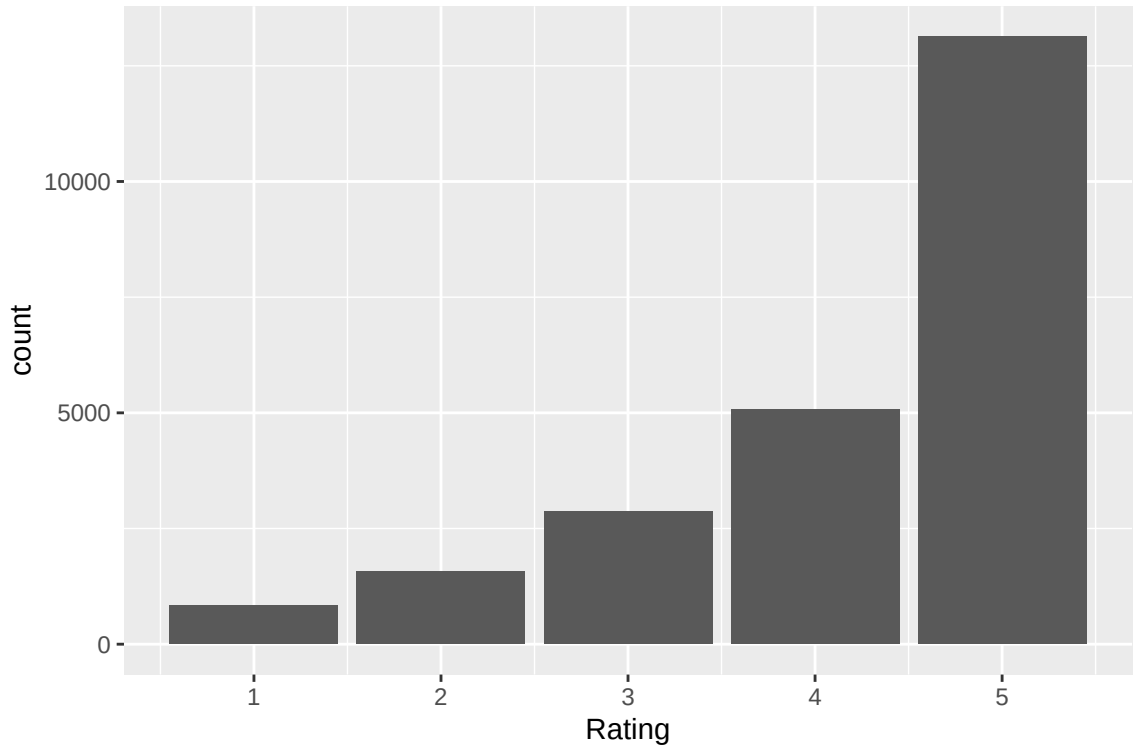
Minimum	Maximum	Average	Median	Standard_Deviation
18	99	43.19854	41	12.27954

From the plot, the ages of the reviewers are skewed to the right. The average age is 43.19854. The maximum age and minimum age are respectively 18 and 99.

Rating

The reviewer rating of the article is on 0-5 scale, so it is a discrete random variable. We could use a bar graph to estimate its distribution.

```
ggplot(review, aes(x=Rating)) + geom_bar()
```




```
review %>% group_by(Rating) %>% summarise(count=n()) %>%
mutate(prop=count/sum(count)) %>% arrange(desc(prop)) %>% kable()
```

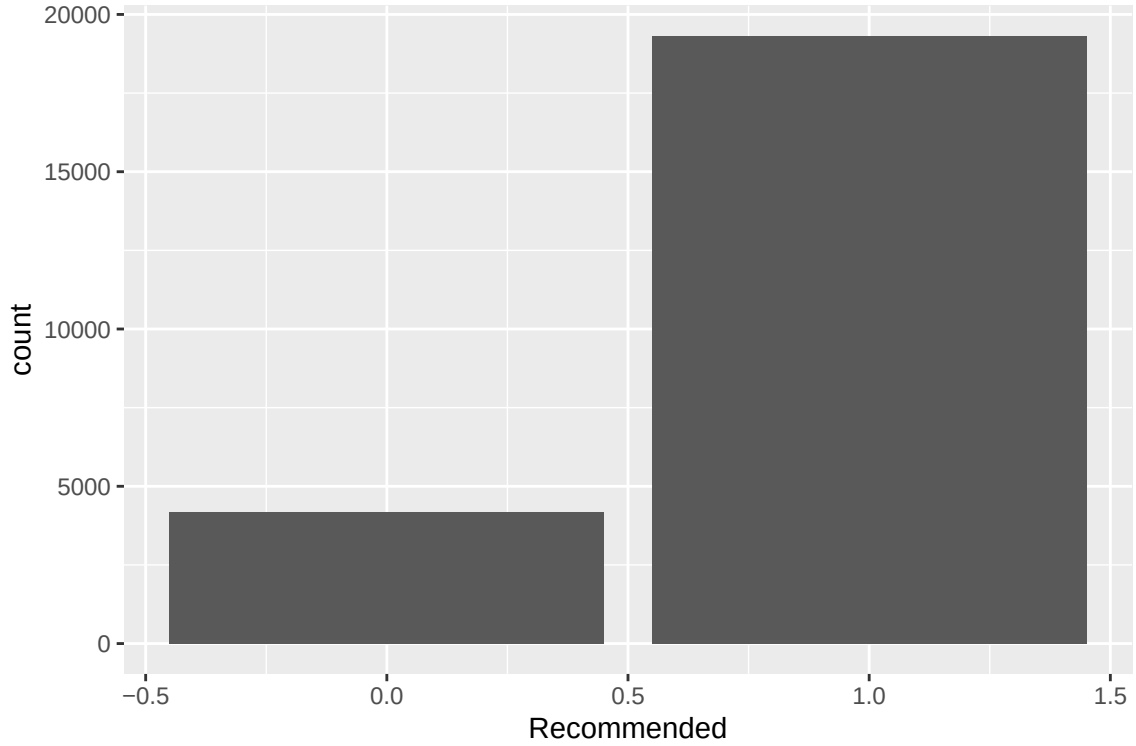
Rating	count	prop
5	13131	0.5590990
4	5077	0.2161713
3	2871	0.1222430
2	1565	0.0666354
1	842	0.0358511

From the bar graph, it's clear that the rating is positively correlated with the number of reviewers. From the numerical summary, 55.9% of the reviewers give a rating of 5, and the number of reviews decreases as the rating decreases.

Recommended

Recommended takes values 0 or 1, so it's also a categorical variable. Similarly, we could use the bar graph.

```
ggplot(review, aes(x=Recommended)) + geom_bar()
```




```
review %>% group_by(Recommended) %>% summarise(count=n()) %>%
mutate(prop=count/sum(count)) %>% arrange(desc(prop)) %>% kable()
```

Recommended	count	prop
1	19314	0.8223623
0	4172	0.1776377

From the bar graph and numerical summary, the majority of the reviewers (82.24%) recommend the products they reviewed. Only 17.76% of the reviewers don't recommend it.

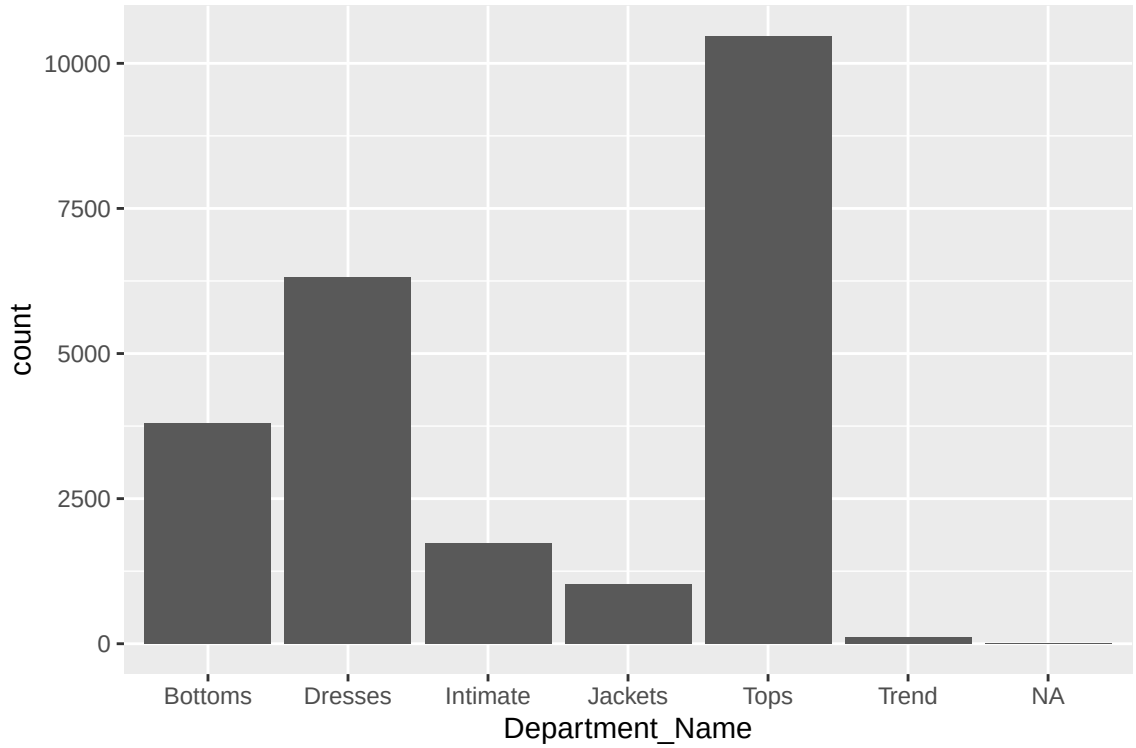
Department_Name

```
unique(review$Department_Name)
```

```
## [1] "Intimate" "Dresses"  "Bottoms"  "Tops"     "Jackets"  "Trend"
## [7] NA
```

There are 6 departments, so Department_Name is also a categorical variable. We also use the bar graph to estimate its distributions.

```
ggplot(review, aes(x=Department_Name)) + geom_bar()
```




```
review %>% 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

From the bar graph, the majority of the reviewers review the clothing from the "Tops" and "Dresses" department. The proportions are respectively, 44.57% and 26.91%. 3799 reviews are from the "Bottoms" department, and the proportion is 16.18%. The sum of the proportions of other departments is around 12%. 14 observations are not available.

We need to remove the missing values from the data-set for the remainder of the tasks.

```
review2 = na.omit(review)
dim(review2)
```

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

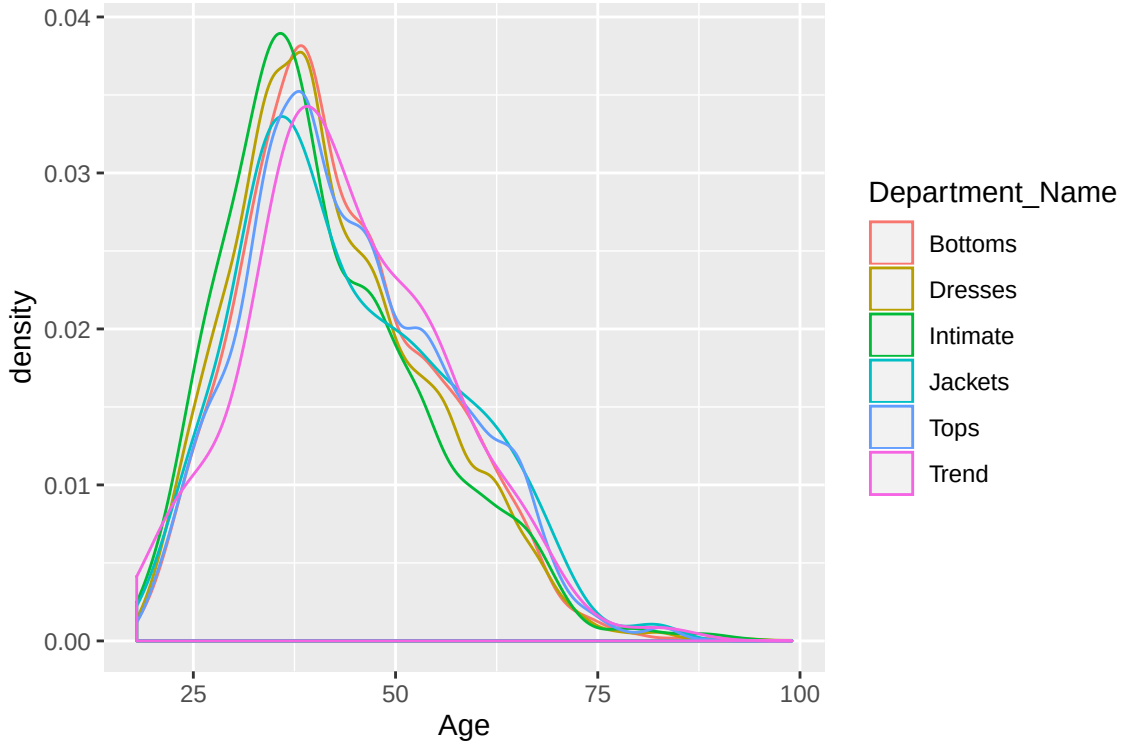
Task 2: Exploring associations

Question 1

The firm would like to know whether the distribution of age of reviewers varies across product departments.

To compare the distribution of age of reviewers across product departments, we should estimate the conditional distributions of age given different product departments. We could estimate the probability density of Age in each department.

```
ggplot(review2,aes(x=Age,col=Department_Name)) +
geom_density(size=0.5)
```



We can also compute the numerical summary of Age for each department.

```
review2 %>% group_by(Department_Name) %>% summarise(Minimum=min(Age),
Median = median(Age), Average = mean(Age),
Maximum = max(Age), Standard_Deviation = sd(Age))
```

```
## # A tibble: 6 x 6
##   Department_Name Minimum Median Average Maximum Standard_Deviation
##   <chr>             <dbl>  <dbl>   <dbl>   <dbl>              <dbl>
## 1 Bottoms              18     41    43.1      92               11.8
## 2 Dresses              18     40    42.1      99               12.0
## 3 Intimate             19     39    41.3      93               12.3
## 4 Jackets              19     42    44.0      83               13.0
## 5 Tops                 18     42    44.1      99               12.5
## 6 Trend                20     43    44.1      83               12.3
```

From the density plot, the distributions of Age in each department are very similar. All the density functions are maximized when Age is around 30, and all the distributions are skewed to the right. From the numerical summary, there is no significant difference between the mean and the standard deviation of age across all departments. All the means are around 42, and all the standard deviations are approximately 12. Thus, we can conclude there is no sufficient evidence that the distribution of age of reviewers varies across product departments.

Question 2

For marketing purposes, they would also like to divide respondent age into five demographic categories: 25 and under, 26 - 35, 36-45, 46-64, and 65 and over and compare the distribution of product ratings amongst each of the five age groups to see which groups are most enthusiastic about their company's products.

```
review3 = review2;
review3 = review3 %>% mutate(Age_Group = "Unknown")
review3$Age_Group[review3$Age <= 25] = "0-25"
review3$Age_Group[review3$Age >= 26 & review3$Age <= 35] = "26-35"
review3$Age_Group[review3$Age >= 36 & review3$Age <= 45] = "36-45"
review3$Age_Group[review3$Age >= 46 & review3$Age <= 64] = "46-64"
review3$Age_Group[review3$Age >= 65] = "65+"
head(review3)
```

```
## # A tibble: 6 x 7
##   Review_ID Clothing_ID   Age Rating Recommended Department_Name Age_Group
##       <dbl>       <dbl> <dbl>  <dbl>       <dbl> <chr>           <chr>
## 1         0         767    33      4           1 Intimate        26-35
## 2         1        1080    34      5           1 Dresses         26-35
## 3         2        1077    60      3           0 Dresses         46-64
## 4         3        1049    50      5           1 Bottoms         46-64
## 5         4         847    47      5           1 Tops            46-64
## 6         5        1080    49      2           0 Dresses         46-64
```

To compare the distribution of product ratings amongst each of the five age groups, we should estimate the conditional distributions of Rating given different age groups. The reviewer rating of the article is on the 0-5 scale, so it is a discrete random variable. We could compute the proportion of each Rating in each age group.

```
prop_table = review3 %>% group_by(Age_Group,Rating) %>% summarise(count=n()) %>%
mutate(prop=count/sum(count))
prop_table$count = NULL
head(prop_table)
```

```
## # A tibble: 6 x 3
## # Groups:   Age_Group [2]
##   Age_Group Rating   prop
##   <chr>      <dbl>  <dbl>
## 1 0-25           1 0.0295
## 2 0-25           2 0.0483
## 3 0-25           3 0.112
## 4 0-25           4 0.221
## 5 0-25           5 0.589
## 6 26-35          1 0.0370
```

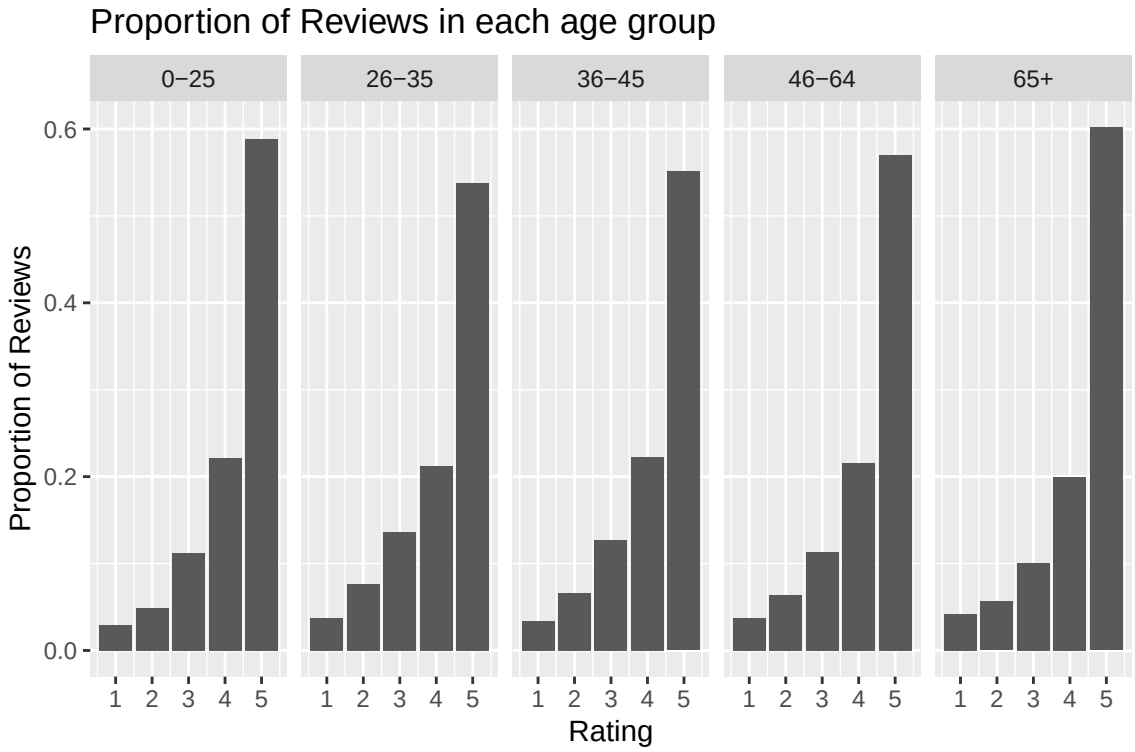
```
prop_table %>% pivot_wider(names_from=Rating,values_from=prop)
```

```
## # A tibble: 5 x 6
## # Groups:   Age_Group [5]
##   Age_Group    `1`    `2`   `3`   `4`   `5`
##   <chr>      <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 0-25      0.0295 0.0483 0.112 0.221 0.589
## 2 26-35     0.0370 0.0769 0.137 0.212 0.538
## 3 36-45     0.0338 0.0662 0.127 0.222 0.551
## 4 46-64     0.0369 0.0643 0.113 0.216 0.570
## 5 65+       0.0418 0.0563 0.100 0.200 0.602
```

From the proportions table, we can see that there is no obvious difference in the conditional distributions of

Rating given different age groups. In all age groups, the proportions increase as the Rating increases, and the proportions are approximately (3%, 6%, 11%, 20%, 60%) for Rating from 1 to 5. We can plot the proportions.

```
ggplot(prop_table, aes(x=Rating, y=prop)) + geom_bar(stat="identity") +
  facet_grid(~Age_Group) + labs(y="Proportion of Reviews",
title="Proportion of Reviews in each age group")
```



From the plot, the distributions of Rating in each age group are very similar. Next, we need to check which groups are most enthusiastic about their company's products.

```
prop_table_love = subset(prop_table,Rating==4 | Rating==5)
prop_table_love %>% group_by(Age_Group) %>% mutate(Prop_High_Rating = sum(prop)) %>%
  arrange(Prop_High_Rating) %>% select(Age_Group,Prop_High_Rating) %>% unique() %>%
  kable()
```

Age_Group	Prop_High_Rating
26-35	0.7494349
36-45	0.7733550
46-64	0.7857143
65+	0.8015873
0-25	0.8099918

From this table, we can see that age groups 65+ and 0-25 are most enthusiastic about their company's products. More than 80% of reviewers from these age groups give a Rating of 4 or 5.

Task 3

For the final task, the company would like to compile a list of their ten most popular products based on recommendations.

```r
##  Compute the number of reviews and the proportion of positive recommendations
##  for every Clothing ID
prop_pos_rating = review2 %>% group_by(Clothing_ID,Recommended) %>% summarise(count=n()) %>%
  mutate(total_count= sum(count), prop=count/sum(count)) %>%
  subset(Recommended == 1 | (Recommended==0 & prop==1)) %>%
  mutate(prop_positive = ifelse(Recommended == 1,prop,0)) %>%
  select(Clothing_ID,total_count,prop_positive)
head(prop_pos_rating)
```

```
## # A tibble: 6 x 3
## # Groups:   Clothing_ID [6]
##   Clothing_ID total_count prop_positive
##         <dbl>       <int>         <dbl>
## 1           0           1         1
## 2           1           3         0.667
## 3           2           1         1
## 4           3           1         1
## 5           4           1         1
## 6           5           1         1
```

```r
## Compute Wilson's lower confidence limit
compute_WLCL = function(pi,ni){
  ai = 1.96*1.96/2/ni
  bi = pi*(1-pi)/ni
  ci = ai/(2*ni)
  WLCL = (pi+ai-1.96*sqrt(bi+ci))/(1+2*ai)
  return(WLCL)
}
```

```r
## Add the WLCL to previous tibble
prop_with_WLCL = prop_pos_rating %>%
  mutate(WLCL=compute_WLCL(prop_positive,total_count))
head(prop_with_WLCL)
```

```
## # A tibble: 6 x 4
## # Groups:   Clothing_ID [6]
##   Clothing_ID total_count prop_positive  WLCL
##         <dbl>       <int>         <dbl> <dbl>
## 1           0           1         1     0.207
## 2           1           3         0.667 0.208
## 3           2           1         1     0.207
## 4           3           1         1     0.207
## 5           4           1         1     0.207
## 6           5           1         1     0.207
```

```r
## Compute the average rating for each Clothing ID
Mean_Rating = review3 %>% group_by(Clothing_ID) %>% select(Clothing_ID,Rating) %>%
  summarise_all(list(Avg_Rating=mean))
head(Mean_Rating)
```

```
## # A tibble: 6 x 2
##   Clothing_ID Avg_Rating
##         <dbl>      <dbl>
## 1           0          5
## 2           1          4
## 3           2          4
```

```
## 4           3          5
## 5           4          5
## 6           5          5
```

```
## Find the Department for each Clothing ID
Dept = review3 %>% select(Clothing_ID,Department_Name) %>% unique() %>%
  arrange(Clothing_ID)
head(Dept)
```

```
## # A tibble: 6 x 2
##   Clothing_ID Department_Name
##         <dbl> <chr>
## 1           0 Jackets
## 2           1 Intimate
## 3           2 Tops
## 4           3 Tops
## 5           4 Tops
## 6           5 Tops
```

```
## Merge all the tibbles together
final_tibble = inner_join(prop_with_WLCL,Mean_Rating,by = "Clothing_ID")
final_tibble = inner_join(final_tibble,Dept,by = "Clothing_ID")
head(final_tibble)
```

```
## # A tibble: 6 x 6
## # Groups:   Clothing_ID [6]
##   Clothing_ID total_count prop_positive  WLCL Avg_Rating Department_Name
##         <dbl>       <int>         <dbl> <dbl>      <dbl> <chr>
## 1           0           1         1     0.207          5 Jackets
## 2           1           3         0.667 0.208          4 Intimate
## 3           2           1         1     0.207          4 Tops
## 4           3           1         1     0.207          5 Tops
## 5           4           1         1     0.207          5 Tops
## 6           5           1         1     0.207          5 Tops
```

List a

```
final_tibble <- final_tibble %>% rename(number_of_review = total_count)
##  the 10 product ID's with the highest average ratings
arrange(final_tibble,desc(Avg_Rating))[1:10,c(1,2,3,5,6)] %>% kable()
```

Clothing_ID	number_of_review	prop_positive	Avg_Rating	Department_Name
0	1	1	5	Jackets
3	1	1	5	Tops
4	1	1	5	Tops
5	1	1	5	Tops
6	1	1	5	Tops
7	1	1	5	Jackets
12	1	1	5	Tops
14	1	1	5	Intimate
16	1	1	5	Dresses
17	1	1	5	Dresses

List b

```
## the 10 product ID's with the highest proportion of positive recommendations
arrange(final_tibble,desc(prop_positive))[1:10,c(1,2,3,5,6)] %>% kable()
```

Clothing_ID	number_of_review	prop_positive	Avg_Rating	Department_Name
0	1	1	5	Jackets
2	1	1	4	Tops
3	1	1	5	Tops
4	1	1	5	Tops
5	1	1	5	Tops
6	1	1	5	Tops
7	1	1	5	Jackets
9	1	1	4	Bottoms
10	1	1	4	Intimate
12	1	1	5	Tops

List c

```
## the 10 product ID's with the highest Wilson lower confidence limits
arrange(final_tibble,desc(WLCL))[1:10,c(1,2,3,5,6)] %>% kable()
```

Clothing_ID	number_of_review	prop_positive	Avg_Rating	Department_Name
1123	30	1.0000000	4.700000	Jackets
834	150	0.9333333	4.540000	Tops
1025	125	0.9360000	4.464000	Bottoms
1008	186	0.9139785	4.462366	Bottoms
984	175	0.9142857	4.462857	Jackets
839	48	0.9583333	4.562500	Tops
1024	35	0.9714286	4.657143	Bottoms
1033	220	0.8954545	4.427273	Bottoms
872	545	0.8770642	4.383486	Tops
1026	21	1.0000000	4.809524	Bottoms

I think the list with 10 product ID's with the highest Wilson lower confidence limits best represents the products which are the most popular.

From List a or List b, all the products have a high average review or a high proportion positively recommended. However, the number of reviews is very small, so there's significant uncertainty about the popularity of these products. The Wilson lower confidence limit is a measure of popularity, which balances both the number of reviews with the proportion recommended, so List c best represents the products that are the most popular.