Data Analysis for Clothing Firm

Zhuoran Zhao

Background and Introduction

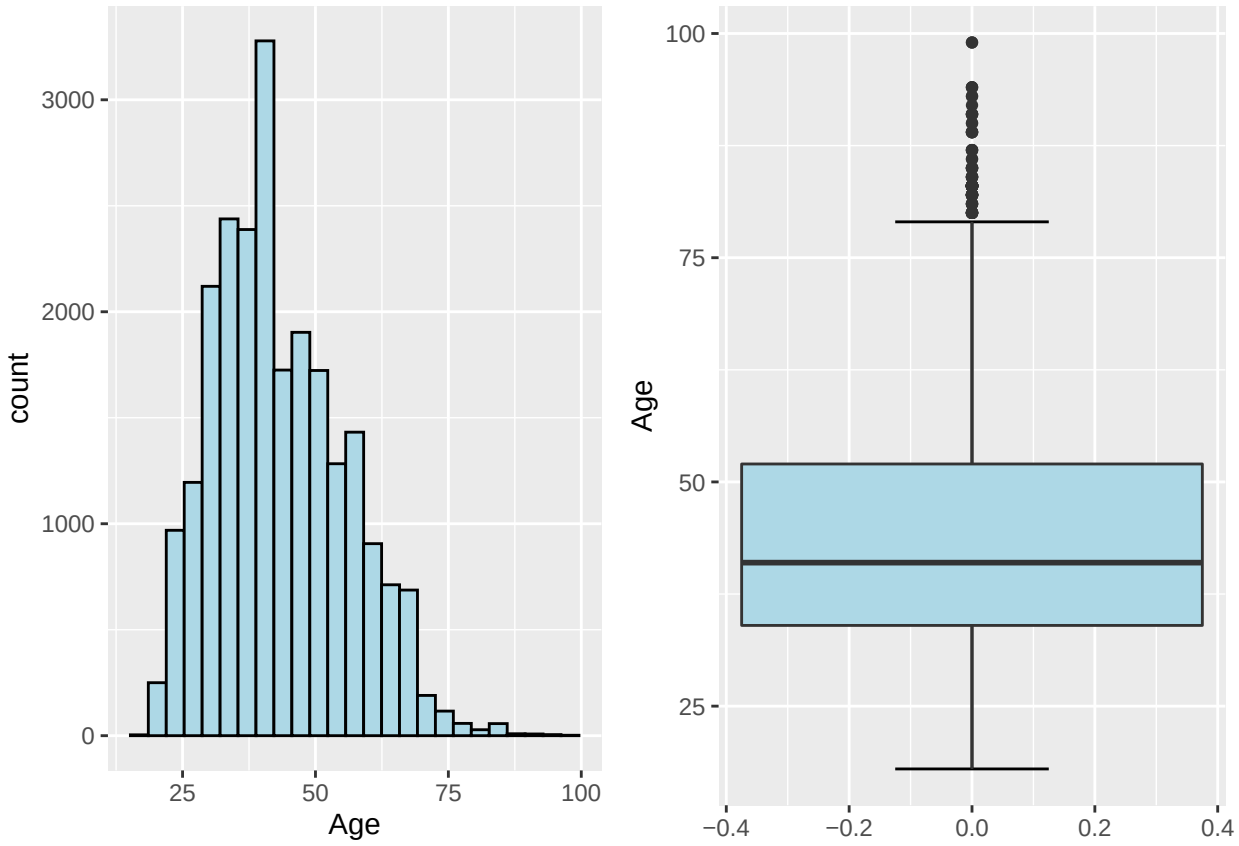
This report provides exploratory analyses through 3 parts for a womens clothing firm. In first section we will discuss about feature of the basic ages, product rating and etc independently from their distributions and basic data summary. Then we will explore the assocations between the age of reviews and product departments. Also, how does the age group and enthusiasm relate the products. Finally, we will depend on the data analysis to recommend a list of 10 most popular products for the company to further consider.

Exploratory Data Analysis

In this section, we will do data analysis via 4 variables that affect the sale. They are age, product rating, recommendations, article departments amongst the respondents. Particularly, we will discuss their distributions and through various analysis among their features.

Age

```
p1 <- ggplot(data, aes(x=Age)) + geom_histogram(bins=25, col='black', fill = 'lightblue')
p2 <- ggplot(data, aes(y=Age)) + stat_boxplot(geom = 'errorbar', width = 0.25) +
      geom_boxplot(fill ='lightblue')
grid.arrange(p1, p2, nrow=1)
```



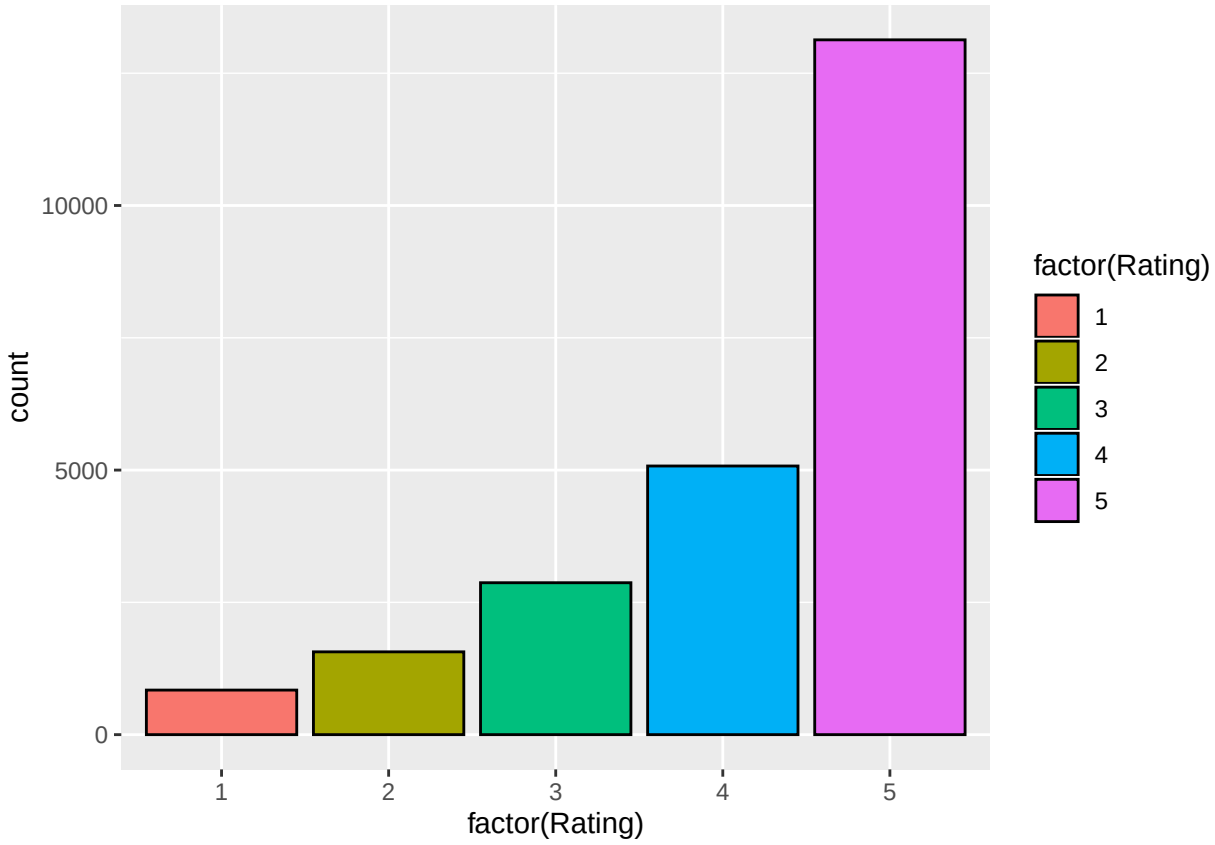

```
data %>%
  select(Age) %>%
  summarise(Age_Mean = mean(Age), Age_Median = median(Age),
            Age_Max = max(Age), Age_Min = min(Age)) %>%
  kable()
```

Age_Mean	Age_Median	Age_Max	Age_Min
43.19854	41	99	18

It can be seen from the table that the range for the customers' ages is pretty wide that from 18 to 99 years old with average around 43 years old. Since their distribution is skewed, that the review number of customers whose age is below 50 tend to be greater than the old ones.

Product Rating

```
ggplot(data, aes(x=factor(Rating), fill=factor(Rating))) +
  geom_bar(col = 'black')
```




```
data %>% group_by(Rating) %>% summarise(n = n()) %>%
  mutate(Proportion  = n/sum(n)) %>% arrange(desc(n)) %>% kable()
```

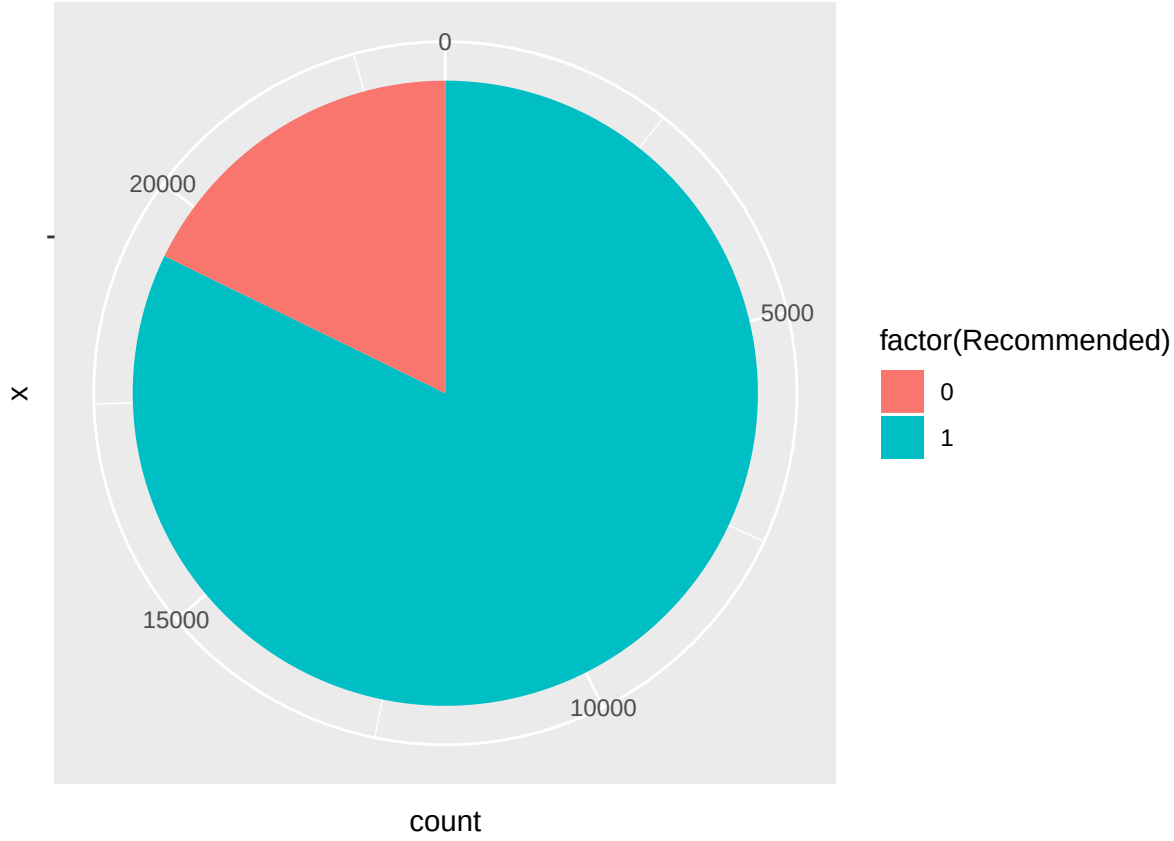
Rating	n	Proportion
5	13131	0.5590990
4	5077	0.2161713
3	2871	0.1222430
2	1565	0.0666354
1	842	0.0358511

It could be seen from the result that the customers tend to give larger rating. And the highest rating 5 has the largest counts and proportion compared to the lowest rating 1. Also, it is noticeable that more than half

of people give a 5-rating(that is around 55 percent).

Recommendations

```
ggplot(data, aes(x="", fill = factor(Recommended))) +
  geom_bar() + coord_polar("y", start = 0)
```




```
data %>% group_by(Recommended) %>% summarise(n = n()) %>%
  mutate(Proportion = n/sum(n)) %>% arrange(desc(n)) %>% kable
```

Recommended	n	Proportion
1	19314	0.8223623
0	4172	0.1776377

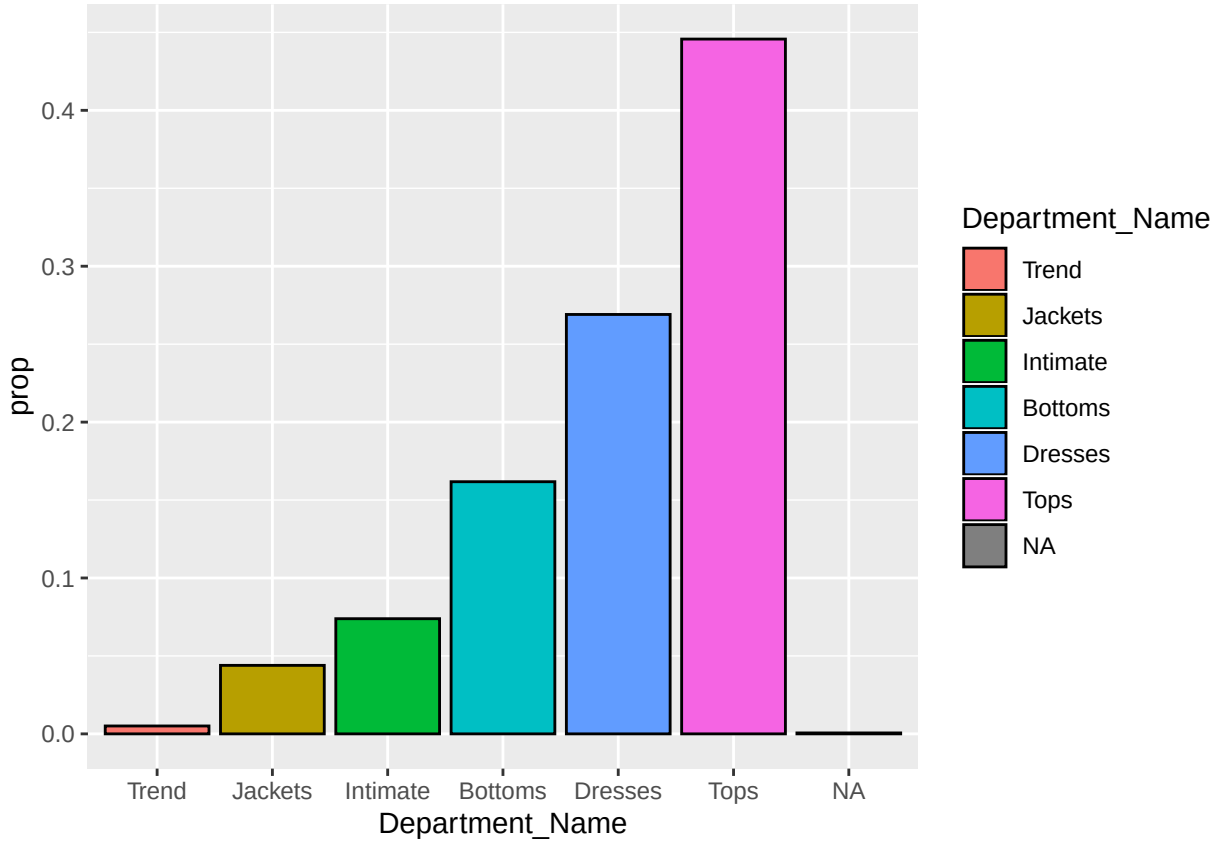
For the recommended data, there are only two possibilities "recommended" or "not recommended". As we could see from both pie chart and summary table that the total proportion of recommended is almost 4 times as those not recommended. It is very obvious that majority of people tend to recommend the product.

Article departments amongst the respondents

```
datamul <- data  %>% group_by(Department_Name) %>%count() %>%
  ungroup() %>% mutate(prop = n/sum(n)) %>% arrange(n)
datamul %>% kable()
```

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

```
data_reorder <- datamul %>% mutate(Department_Name = fct_reorder(Department_Name, n))
ggplot(data_reorder, aes(x=Department_Name, fill=Department_Name)) +
  geom_bar(stat='identity', aes(y=prop),col='black')
```



We could see that the largest amout of reviews is in Tops departments, followed by Dresses and Bottoms. The least popular department is Trend. There also a small amout of reviews that do not include department names.

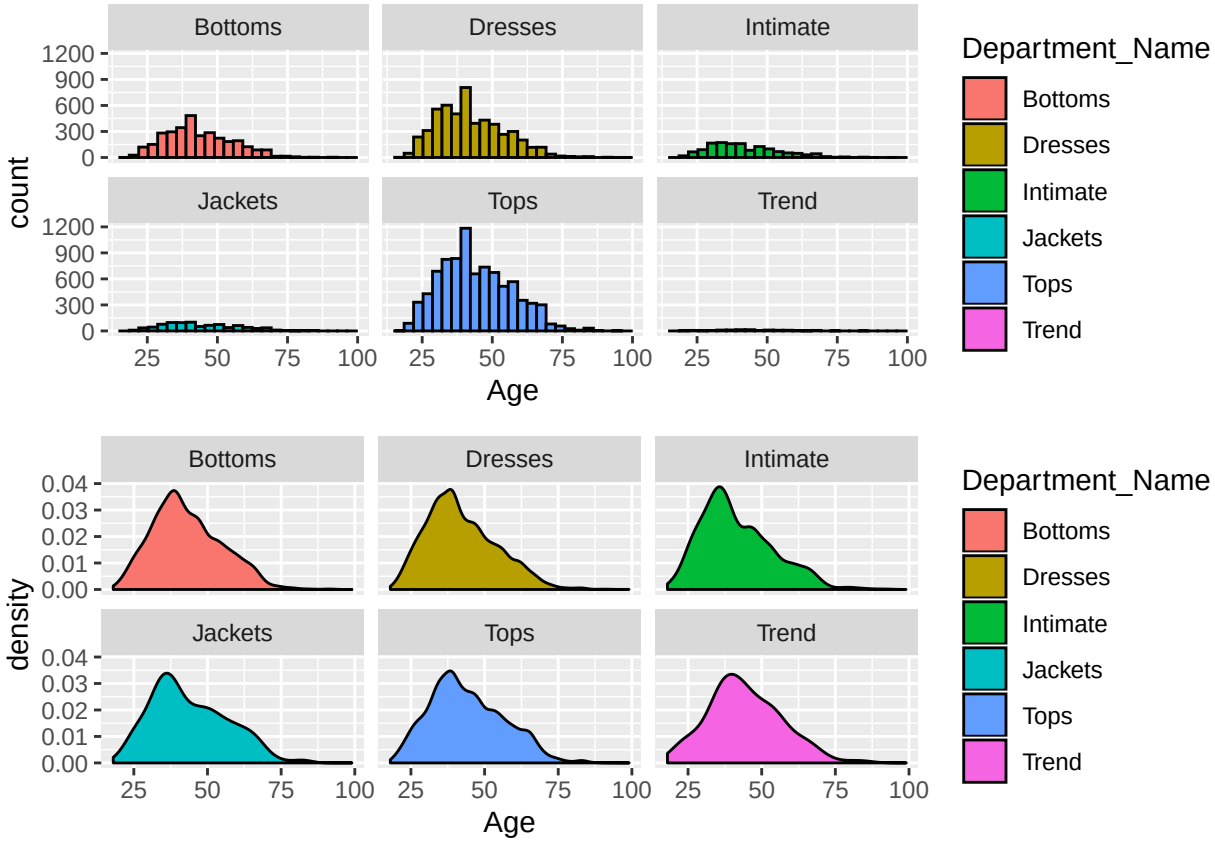
Exploratory associations

Distribution of age of reviewers vs. product departments

```
data <- na.omit(data)
p1 <- ggplot(data, aes(x=Age, fill = Department_Name), group = Department_Name) +
  geom_histogram(bins = 25, col='black') + facet_wrap(~Department_Name)

p2 <- ggplot(data, aes(x=Age, fill = Department_Name), group = Department_Name) +
  geom_density() + facet_wrap(~Department_Name)


grid.arrange(p1, p2)
```

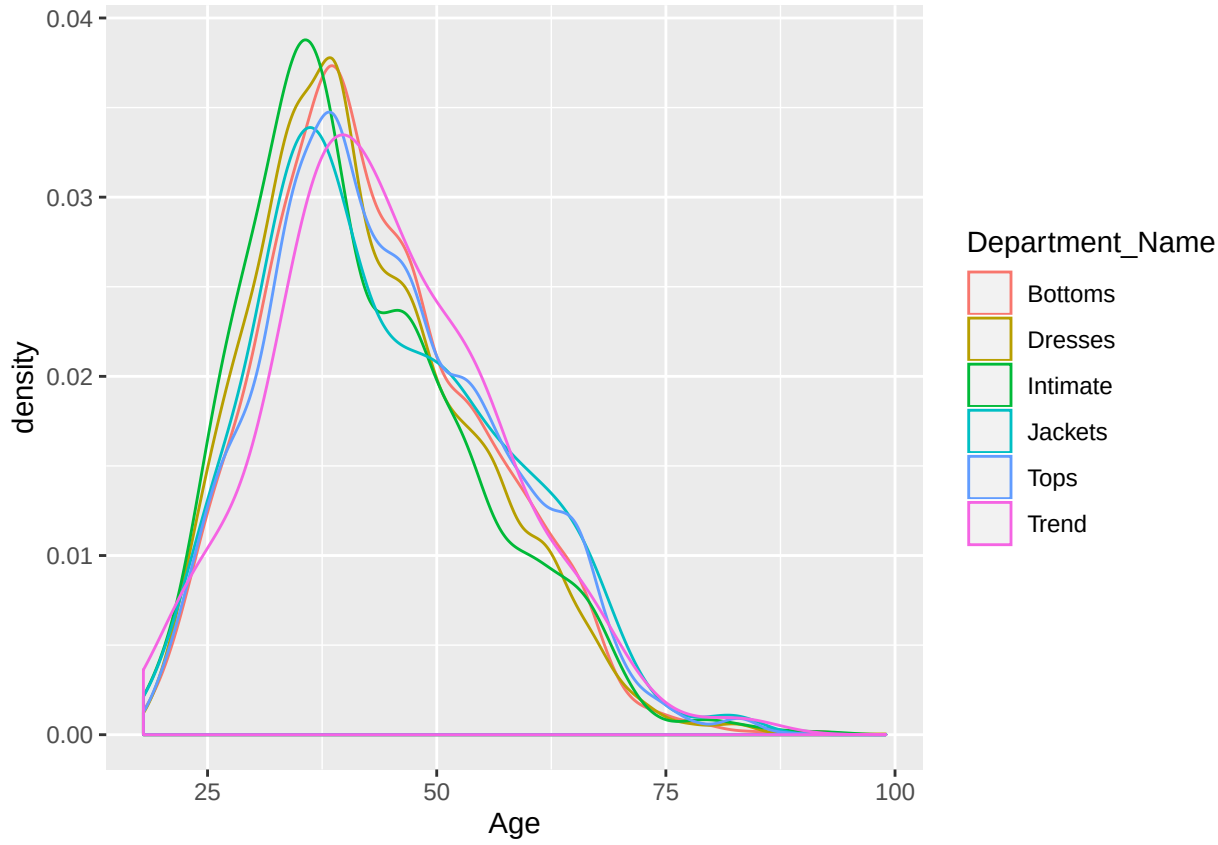



```
data %>% group_by(Department_Name) %>% summarise(Age_Mean=mean(Age), Age_Median=median(Age),
Age_Min=min(Age), Age_Max=max(Age)) %>% kable()
```

Department_Name	Age_Mean	Age_Median	Age_Min	Age_Max
Bottoms	43.18467	41	18	92
Dresses	42.19307	40	18	99
Intimate	41.63352	39	19	93
Jackets	43.96132	42	19	83
Tops	44.12579	42	18	93
Trend	44.34579	43	20	83

From the histogram we could see that the total amount of purchase for each department is quite different. However, when comparing their densities, they are very similar to each group. Also, we could find the similar trend from the table. Among all different departments, they all have very closed average, median, min and max ages.

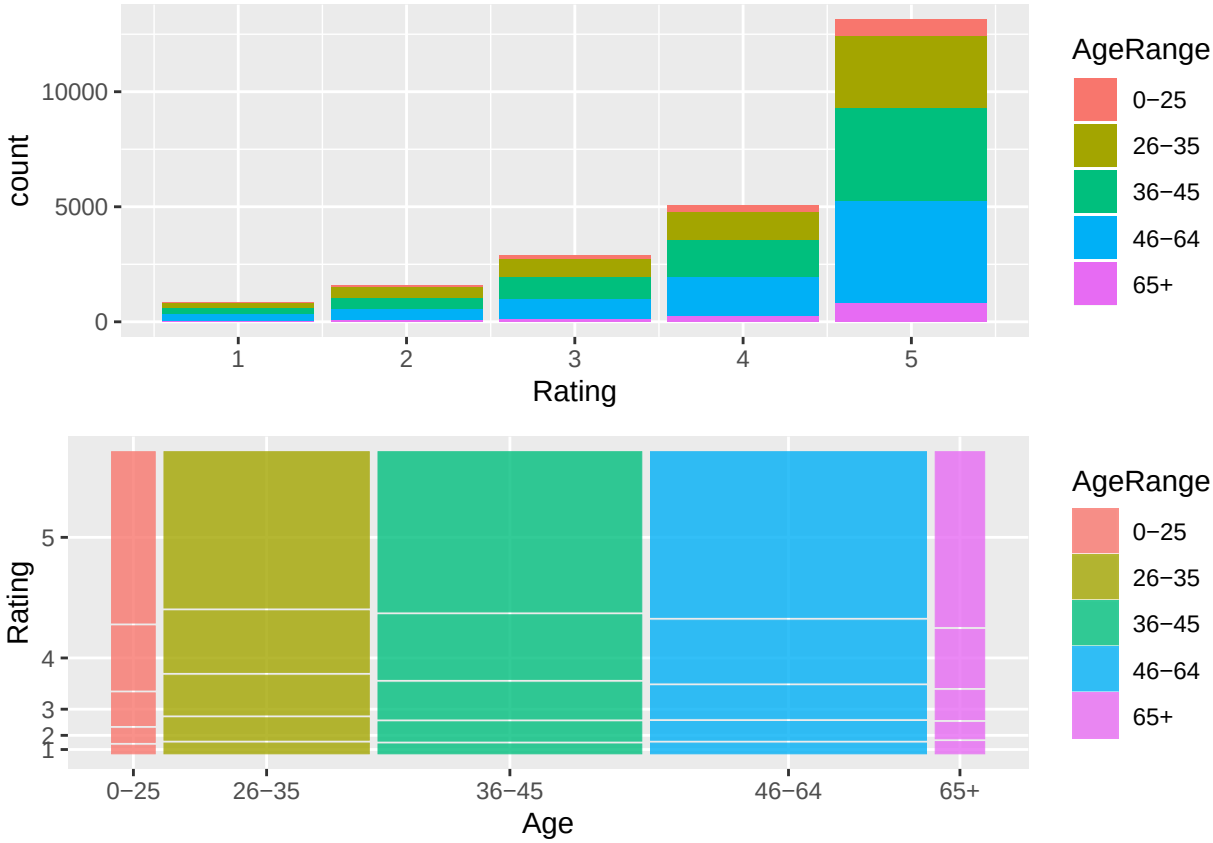
```
ggplot(data, aes(x=Age, col=Department_Name)) + geom_density()
```



Further, in order to prove our assumption, we put all group's density in one graph. The overlap between different group better prove our conclusion. Then we could conclude that the distribution of age of reviewers do not vary across product departments.

Enthusiastic Age Range

```
data <- data %>% mutate(AgeRange ='')
data$AgeRange[data$Age <=25] <- "0-25"
data$AgeRange[data$Age >=26 & data$Age<=35] <- "26-35"
data$AgeRange[data$Age >=36 & data$Age<=45] <- "36-45"
data$AgeRange[data$Age >=46 & data$Age<=64] <- "46-64"
data$AgeRange[data$Age >=65] <- "65+"
p1 <- ggplot(data, aes(x=Rating, fill = AgeRange)) + geom_bar()
p2 <- ggplot(data) + geom_mosaic(aes(x=product(Rating, AgeRange),
                                     fill=AgeRange)) + xlab("Age") + ylab("Rating")
grid.arrange(p1, p2)
```



We divided our dataset into five group with age range indicated in the graph.Through the stacked bar plot we could see that the largest amount of 5-rating comes from age group 46-64. However, when we see the mosaic plot, we could see that although the total amount of reviews in 0-25 group is the smallest. The relative proportion for them to give a higher rating is the highest.(It has the lowest proportion for rating-1 and highest proportion for rating 5). Thus we would conclude although there is no significant difference, that age range 0-25 tends to more enthusiastic than others.

10 most popular products

In this section, we would provide the list of 10 most popular products depend on three different measurements.

10 popular products based on highest average ratings

```
data %>% group_by(Clothing_ID) %>%
  summarise(Avg_Rating = mean(Rating)) %>% arrange(desc(Avg_Rating)) %>%
  slice(1:10) %>% kable()
```

Clothing_ID	Avg_Rating
0	5
3	5
4	5
5	5
6	5
7	5
12	5
14	5
16	5
17	5

This list only listed the first 10 highest average ratings, however, there are more than 10 Clothes have rating 5 that did not display.

10 popular products based on highest proportion of positive recommendations

```
result <- data %>% group_by(Clothing_ID) %>%
  mutate(n = n(),Positive_Prop=sum(Recommended)/n) %>% arrange(desc(Positive_Prop))

result %>% select(Clothing_ID, Positive_Prop) %>% unique() %>% head(10) %>% kable()
```

Clothing_ID	Positive_Prop
767	1
1120	1
684	1
4	1
89	1
1196	1
126	1
670	1
329	1
596	1

As the table indicated, they are the 10 popular products that have highest positive proportion. However, since there are more than 10 groups have positive proportion 1, we only display the first 10 depend on their

cloth ID.

10 popular products with highest Wilson lower confidence limits for positive recommendations

```
Wilson_value <- function(p,n){
  a <- 1.96^2/(2*n)
  b <- (p*(1-p))/n
  c <- a/(2*n)
  result <- (p + a - 1.96*sqrt(b+c))/(1+2*a)
}

result_n <- result %>% mutate(Wilson_Value = Wilson_value(Positive_Prop, n)) %>%
  arrange(desc(Wilson_Value))
result_n %>% select(Clothing_ID, Wilson_Value) %>% unique() %>% head(10) %>% kable()
```

Clothing_ID	Wilson_Value
1123	0.8864829
834	0.8816365
1025	0.8787832
1008	0.8648440
984	0.8634038
839	0.8602409
1024	0.8546659
1033	0.8480142
872	0.8468267
1026	0.8453562

We would like to compile a list for company of their 10 most popular products based on recommendations. In order to balance both number of reviews and the proportion recommended, we use the Wilson's lower confidence limit approximation for proportions. As the table indicated the top 10 products depend on the value of Wilson lower confidence limt.

Through these 3 different approaches, I would recommend the Wilson's lower confidence limit approximation measurement. Since it combine both proportion, and the number of respondents, instead of using one of these factors independently. Besides, within the results, we could actually select 10 products, not like first 2 approaches that we may have a list that contains more than 10 products.

Conclusion

Throughout this report, we did exploratory analysis on independent variables, associations. Then we give the 10 most popular products depend on 3 measurements. In the end, we conclude that Wilson's lower condfidence limit approximation measurement gives us the better result. Notice, in this project we only discuss these features through the data analysis, but there are far more techniques such as fitting models, predictions and we could even apply machine learning techniques for further evaluations.

```
x <-tibble(Avg_SD = c(0.6505803), Max_SD = c(1.564442), Min_SD = c(0.0222924))
x %>% kable()
```

Avg_SD	Max_SD	Min_SD
0.6505803	1.564442	0.0222924

```
y <- tibble(Avg_DU = c(0.0158095), Max_DU = c(0.1075734), Min_DU = c(-0.0259051))
y %>% kable()
```

Avg_DU	Max_DU	Min_DU
0.0158095	0.1075734	-0.0259051