HW4-Q3

luhuan wu

11/2/2018

```
library(tidyverse)
```

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

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

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

```
# Load data
amazon = read.csv("DatafinitiElectronicsProductData.csv", header=TRUE)
```

`amazon` is a dataset containing reviews for electronic products sold in Amazon, Walmart, BestBuy and other online platforms.

Each row contains one review information by a customer, which is described by 27 variables. There are 7299 reviews in total. Furthermore, the reviews are for 50 different products, which come from 38 different brands.

In this homework, I will conduct analysis on `reviews.rating` variable, which is a discrete variable ranging from 1 to 5. The higher `review.rating` is, the better the customer think the product is. `NA` value indicates that the customer did not rate the product.

1) Missing pattern in `reviews.rating`.

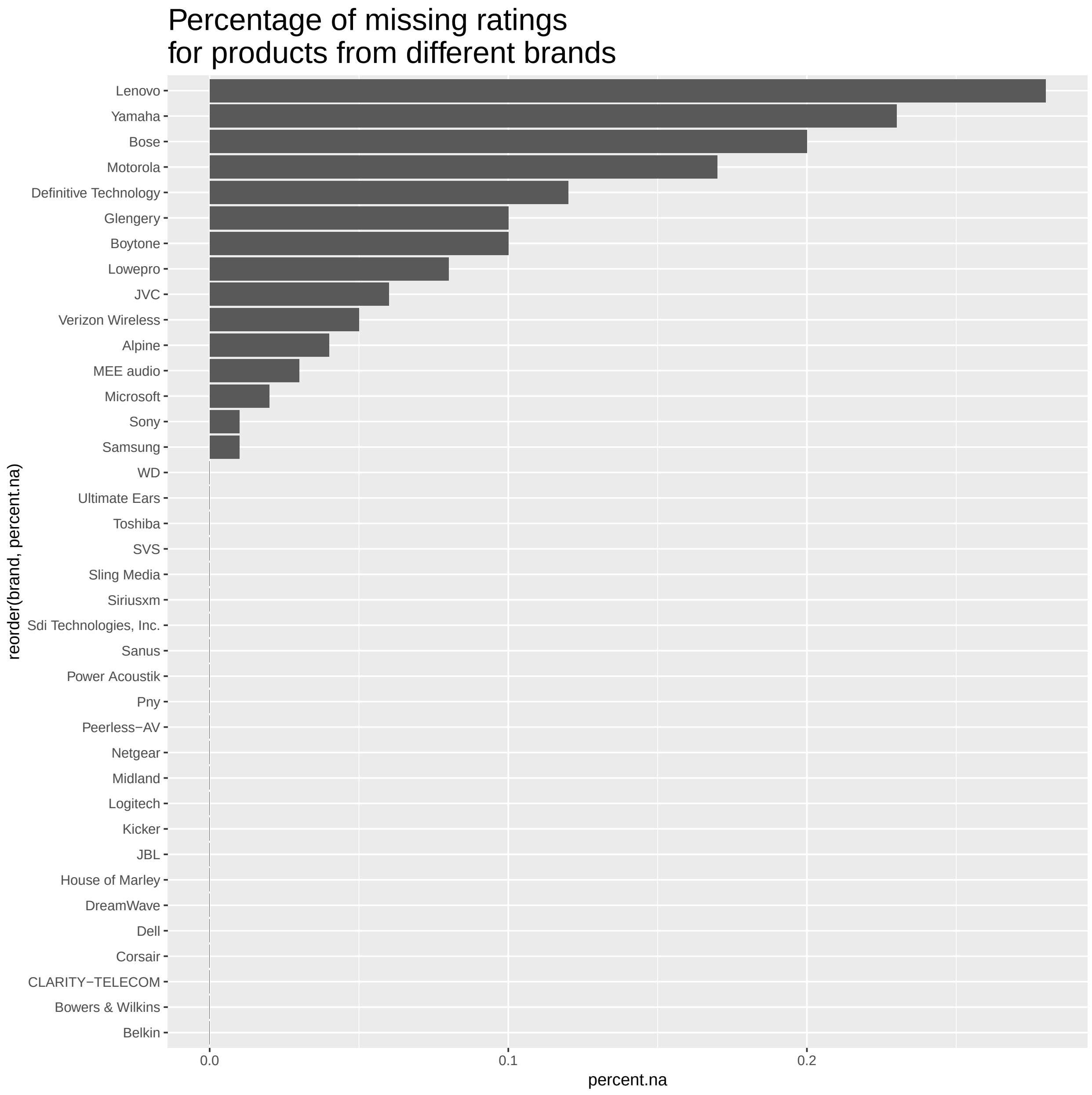
First, we would compute the number and percentage `NA` value of each product.

```
rating.na.df = amazon %>%
  group_by(brand) %>%
  summarise(num.brand = n(), num.na = sum(is.na(reviews.rating))) %>%
  mutate(percent.na = round(num.na / num.brand, 2))%>%
  arrange(desc(percent.na))
rating.na.df
```

```
## # A tibble: 38 x 4
##    brand                  num.brand num.na percent.na
##    <fct>                      <int>  <int>      <dbl>
##  1 Lenovo                        39     11       0.28
##  2 Yamaha                       150     35       0.23
##  3 Bose                          25      5       0.2
##  4 Motorola                      48      8       0.17
##  5 Definitive Technology        123     15       0.12
##  6 Boytone                       83      8       0.1
##  7 Glengery                     127     13       0.1
##  8 Lowepro                      143     12       0.08
##  9 JVC                          142      8       0.06
## 10 Verizon Wireless             148      8       0.05
## # ... with 28 more rows
```

Now, let's see distribution of missing percentage over different brand's products.

```
ggplot(rating.na.df, aes(x=reorder(brand, percent.na), y=percent.na)) +
  geom_bar(stat='identity') +
  coord_flip() +
  ggtitle("Percentage of missing ratings\nfor products from different brands") +
  theme(plot.title = element_text(size=20))
```



From the bar chart above, we could find that Lenovo's producsts are given least ratings in terms of percentage, then follows by Yamaha and Bose. However, the missing percentage are all less than 30%. On the other hand, for those brands that do not have missing reviews, it is likely that the total reviews given are a very small number, such as brand `SVS`, which only has one review.

Now, we first filter out the reviews of the brands whose total reviews are less than 20, and then filter out the reviews that miss ratings.

```r
brands.filter = rating.na.df %>%
  filter(num.brand < 21)

amazon.new = amazon %>%
    filter(! brand %in% brands.filter$brand) %>%
    filter(! is.na(reviews.rating))
```

2) Distribution of `review.rating`

In this part, we will display stacked bar charts of ratings for each product.

First, we calculate the percentage of different levels of `review.rating` for different brands.

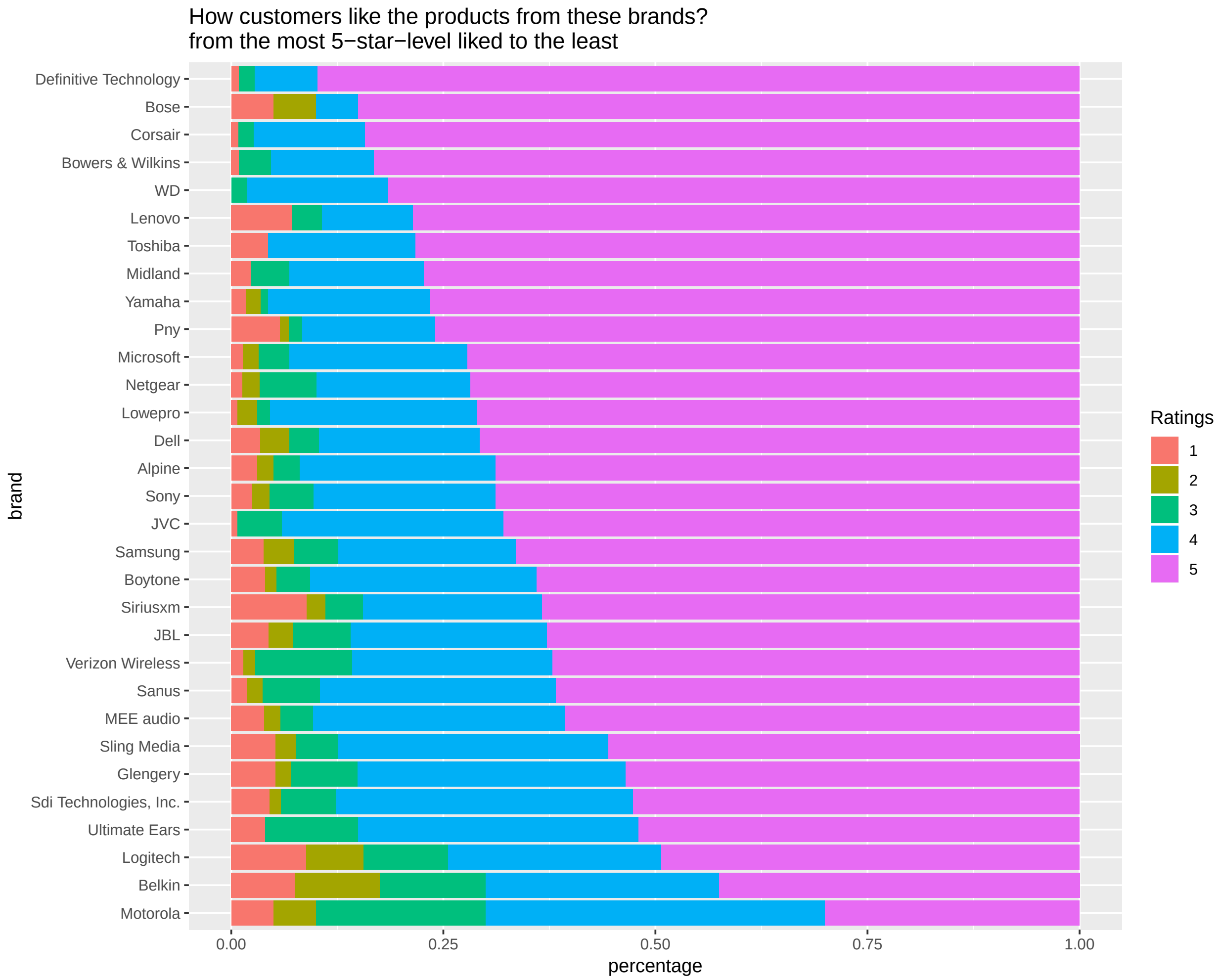
```r
ratings = amazon.new %>%
  group_by(brand, reviews.rating) %>%
  summarise(n = n()) %>%
  transmute(reviews.rating, freq = n / sum(n))

ratings$reviews.rating = factor(ratings$reviews.rating)

ratings$freq5 = 2
for( i in 1:nrow(ratings)){
  bbrand = ratings[i,]$brand
  l = ratings %>% filter(brand==bbrand, reviews.rating==5)
  ratings[i,]$freq5 = l$freq
}
```

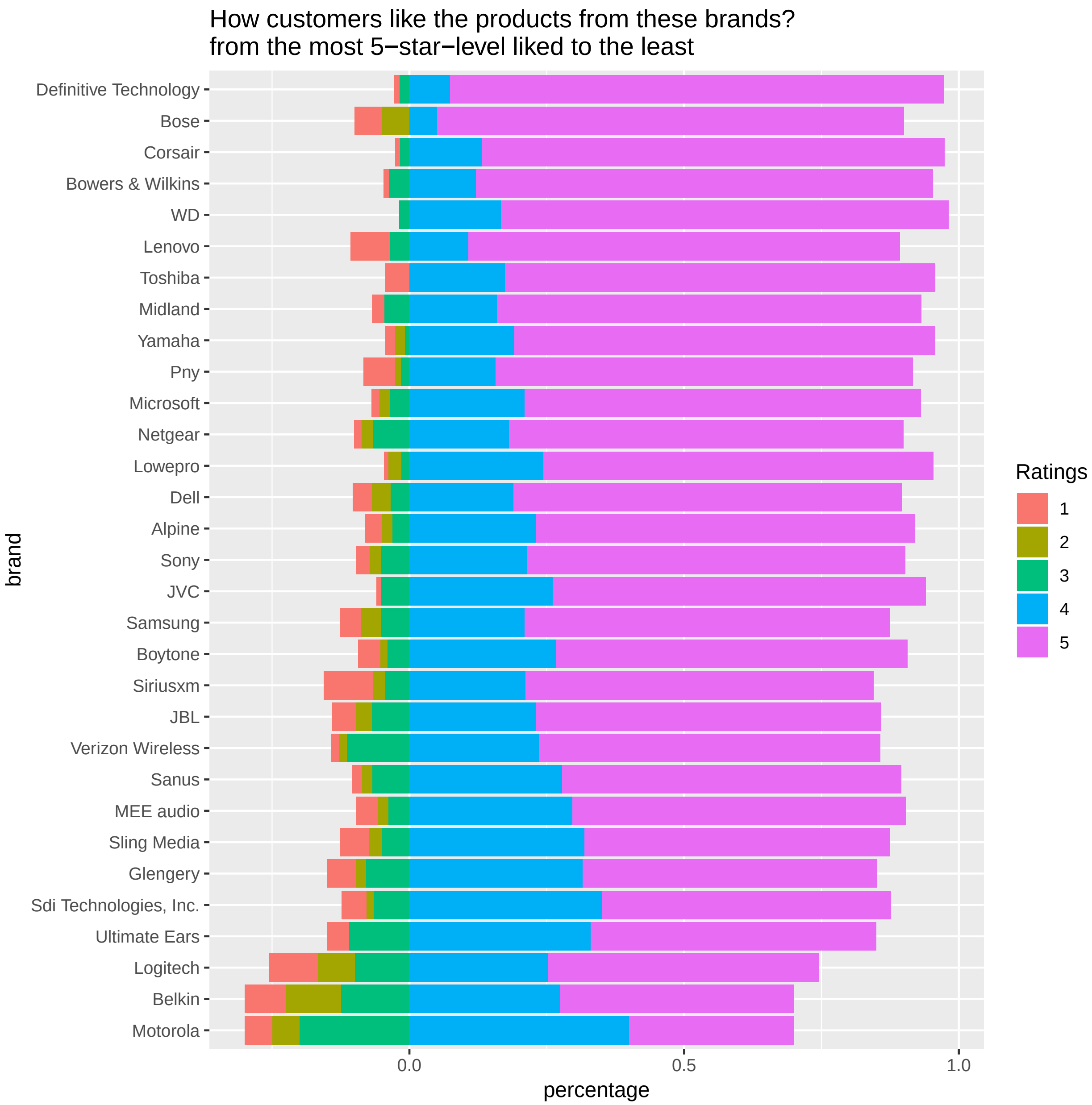
Then, we display the distribution of ratings over brands in descending order of the percentage of 5-star rating.

```r
ggplot(ratings, aes(x = reorder(brand, freq5), y = freq, fill = reviews.rating)) +
  geom_bar(stat = "identity", position = position_fill(reverse = TRUE)) +
  scale_fill_discrete(name = "Ratings") +
  xlab('brand') + ylab('percentage') +
  ggtitle('How customers like the products from these brands?\nfrom the most 5-star-level liked to the l
  coord_flip()
```



We could also display a diverging stacked bar chart, where the midpoint is the division between the 3-star and the 4-star rating. Note that the brands are also displayed in descending order of percentage of 5-star ratings.

```
ggplot(ratings, aes(x = reorder(brand, freq5), fill=reviews.rating)) +
      geom_bar(data = subset(ratings, reviews.rating %in% c(1,2, 3)),
               aes(y = -freq), position="stack", stat="identity") +
      geom_bar(data = subset(ratings,
                             reviews.rating %in% c(4,5)),
               aes(y = freq),
               position = position_stack(reverse = TRUE), stat="identity") +
      xlab('brand') + ylab('percentage') + scale_fill_discrete(name = "Ratings") +
      ggtitle('How customers like the products from these brands?\nfrom the most 5-star-level liked to
      coord_flip()
```



A few thoughts on the diverging bar charts:

- From top to down, as the percentage of 5-star ratings are decreasing (which is due to the plot rule we define), the 'center' of the bar seems to shifted to the left, which suggests the overall rating of the brands are shifting to negative.
- It is naturally to think about plotting the average rating on the diverging bar chart, and see if the average point is shifter to the left as well.
- Customers seems to be conservative about giving an exreamely negative rating, which is 1-star. For those brands that have really low percentage of 5-star ratings, the percentage of 1-star ratings are not relatively high compare to other brands. However, the percentage of 3-star ratings increases as that of 5-star decrease, which could be an interesting take-away.
- For future exploration, we could investigate the relationship of the percentage of 5-star rating and taht of 3-star rating.

- Another direction is to facet the ratings of a brand by different products.

3) Average ratings

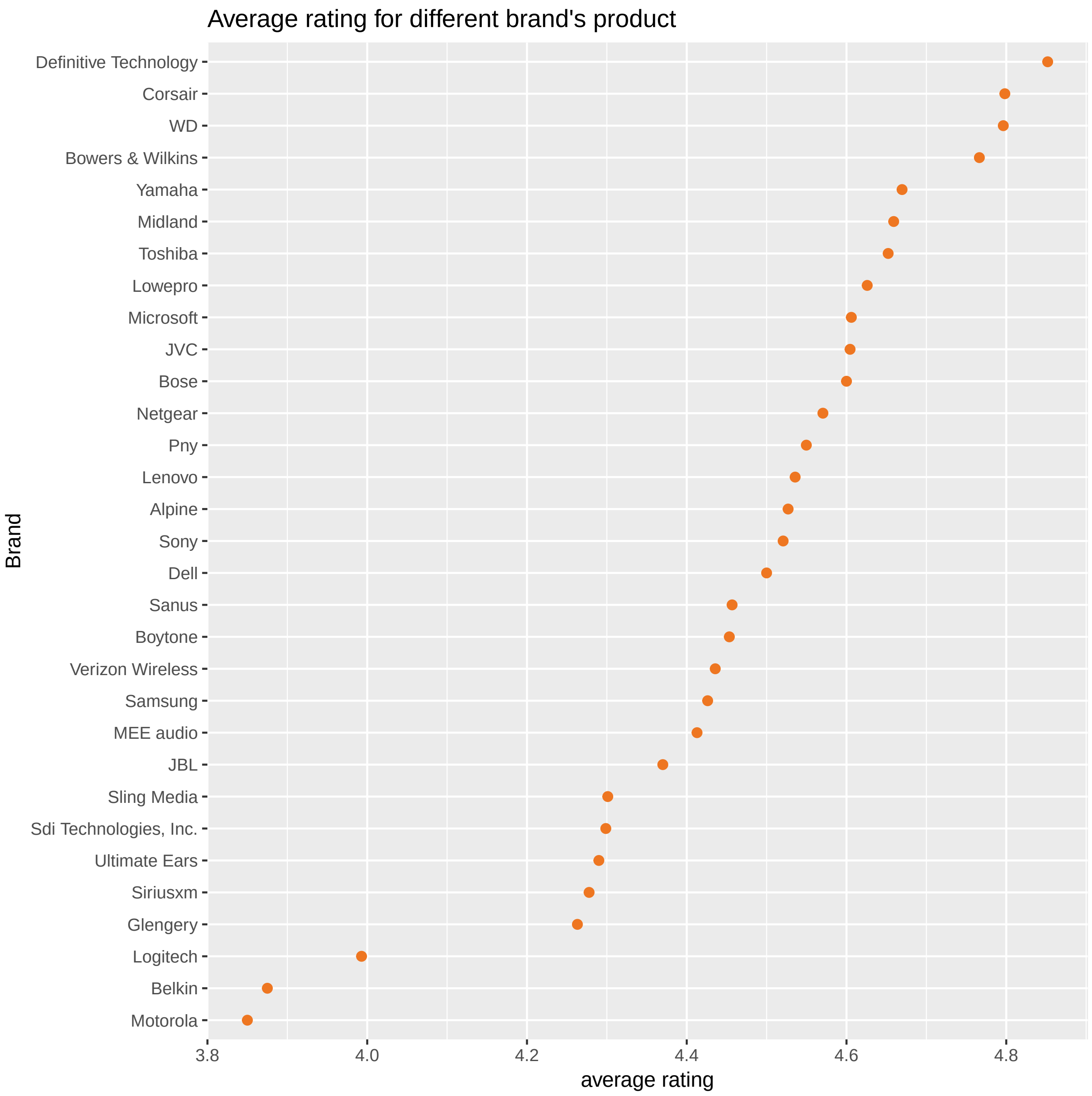
In previous part, we focus on distribution of different levels of ratings. In this part, we will use the average rating as an indicator for positiveness / negativeness of reviews.

```
ratings2 = amazon.new %>%
  group_by(brand, reviews.rating) %>%
  summarise(n = n()) %>%
  transmute(part.sum.rating = reviews.rating * n / sum(n))
```

```
avg.rating =
  aggregate(ratings2$part.sum.rating, by=list(Category=ratings2$brand), FUN=sum)
```

We present a Cleveland dot plot as follows:

```
# create dot plot theme
ggplot(avg.rating, aes(x = x, y = fct_reorder(Category, x))) +
    geom_point(color = "chocolate2", size=2) +
    ylab("Brand") + xlab("average rating") +
    ggtitle("Average rating for different brand's product")
```



We could find that the brands that are of highest ratings have overlaps with the brands of highest percentage of 5-star ratings. Hence, in the future step we will explore the relationship between the two criteria: the percentage of 5-star ratings and average ratings.