Project 1 - Blush Stats from Sephora and ULTA

Hao Cui hc3498

1. Intro

A topic brought up between my friends was how there's such a plethora of beauty products available it seems hard to decide on which to buy. Hence, this project aims to perhaps give clarity on purchasing choices using statstics. I made my own 2 datasets by recording data observed from the online beauty stores Sephora.com and ULTA.com. I chose blush as my focused product (found in both stores). On the websites, each blush product has its own price, five-star review, and number of colors available which serve as my numeric variables and what type of blush (there are three types: cream, liquid, and powder) which serves as my categorical variable. Each observation is a single blush product and 25 were collected from each store that will merge for a total of 50 (there are more than 50 for both stores, so this is just a sampled list). Just scanning through, I don't think there are any particular trends, but we shall see! My sources are from: https://www.sephora.com/shop/blush (https://www.sephora.com/shop/blush) and https://www.ulta.com/makeup-face-blush?N=277v (https://www.ulta.com/makeup-face-blush?N=277v)

Importing Data:

The tidyverse and readxl packages were installed and the two excel sheet datasets were imported with the following code.

```
library(tidyverse)
##install.packages("readxl") in console
library(readxl)

## Import Code had this below:
Sephora_data <- read_excel("C:\\Users\\Hao\\Documents\\Elements of Data Science\\Sephora data.xl
sx")
View(Sephora_data)

ULTA_data <- read_excel("C:\\Users\\Hao\\Documents\\Elements of Data Science\\ULTA data.xlsx")
View(ULTA_data)
```

2. Joining/Merging

Below shows the code used to merge the Sephora and ULTA datasets. Since the variables for the two datasets are the same, full_join was used and the data is already tidy. (Tidying will be used later). As the datasets are essentially the same, having the same six variable types, with just unique observations for each, a full join was used to keep all the x and y observations into a single dataset called "Blush_data". By default R studio joins by the all the variables ("Name", "Brand", "Price", "Review", "Colors", "Type").

```
## install.packages("dplyr") in console
library(dplyr)
Blush_data <- full_join(Sephora_data, ULTA_data)
```

3. Wrangling

Below shows the use of dplyr functions in data wrangling to find out certain things about our data.

Which Brands have the highest review for cream blushes?

```
Blush_data %>%
  filter(Type == "Cream")%>%
  select(-Colors)%>%
  arrange(desc(Review))
```

```
## # A tibble: 19 x 5
##    Name                                        Brand          Price Review Type
##    <chr>                                       <chr>          <dbl>  <dbl> <chr>
##  1 Higher Standard Satin Matte Cream Blush     LYS Beauty     16       4.8 Cream
##  2 Stick Cream Blush                           Anastasia Bev~ 32       4.7 Cream
##  3 Blush Divine Clean Dewy Cream Blush         Rose Inc       30       4.7 Cream
##  4 Cheek Kiss Cream Blush                      Milani          9.99    4.7 Cream
##  5 Soft Pop Blush Stick                        Makeup by Mar~ 28       4.6 Cream
##  6 Chubby Stick Cheek Colour Balm Blush        Clinique       26.5     4.6 Cream
##  7 Pot Rouge for Lips & Cheeks                 Bobbi Brown    34       4.6 Cream
##  8 BeachPlease Lip + Cheek Cream Blush         Tower 28 Beau~ 20       4.5 Cream
##  9 Flush Balm Cream Blush                      Merit          28       4.5 Cream
## 10 Nudies Matte Blush & Bronze                 Nudestix       34       4.5 Cream
## 11 Convertible Color                           Stila          25       4.5 Cream
## 12 Putty Blush                                 e.l.f. Cosmet~  6       4.4 Cream
## 13 Blush Stix                                  ColourPop       9       4.4 Cream
## 14 Stay Vulnerable Melting Cream Blush         Rare Beauty b~ 21       4.3 Cream
## 15 Cheeks Out Freestyle Cream Blush            Fenty Beauty ~ 22       4.3 Cream
## 16 The Multiple Cream Blush, Lip and Eye Stick Nars           39       4.3 Cream
## 17 Multi-Stick Cheek & Lip                     ILIA           34       4.2 Cream
## 18 Lip + Cheek Cream Blush Stick               Milk Makeup    20       4.1 Cream
## 19 Monochromatic Multi Stick                   e.l.f. Cosmet~  4       3.9 Cream
```

The above table looks for which cream blushes have the highest reviews. The data was filtered for just cream blushes, colors column was removed since I figured it was not important in comparison to, say, price for example. The observation was arranged in descending order from the highest review. Brands like LYS beauty, Anastasia Beverly Hills, Rose Inc, and Milani have the highest reviewed cream blushes.

Adding a new variable. Do the number of colors have an effect on price?

```
Blush_data %>%
  mutate(price_per_color = Price/Colors)%>%
  select(-Name, -Brand, -Review, -Type)
```

```
## # A tibble: 50 x 3
##    Price Colors price_per_color
##    <dbl>  <dbl>           <dbl>
##  1    20     11            1.82
##  2    40     23            1.74
##  3    21      5            4.2
##  4    30      3           10
##  5    20      8            2.5
##  6    24      4            6
##  7    20      6            3.33
##  8    22     10            2.2
##  9    24      5            4.8
## 10    29     12            2.42
## # ... with 40 more rows
```

A new variable called price_per_color, calculates the price of the product per number of colors are available. This was done because some say that as the number of colors increase, this makes the product more costly to make, so a higher price. But from the looks of this table that isn't necessarily the case, many expensive products have low colors and vice versa. The business model could be a reason why, where some brands have more investors, which allow them to offer more colors at cheaper prices.

Numerical & Categorical Summary Stats

```
## Numeric
Blush_data %>%
  summarize(mean_review = mean(Review, na.rm = T),
  sd_review = sd(Review), median_price = median(Price))
```

```
## # A tibble: 1 x 3
##   mean_review sd_review median_price
##         <dbl>     <dbl>        <dbl>
## 1        4.44     0.247         24.5
```

```
## Categorical
Blush_data %>%
  mutate(budget_friendly = Price< 12)%>%
  group_by(budget_friendly, Type)%>%
  summarize(count = n())
```

```
## # A tibble: 6 x 3
## # Groups:   budget_friendly [2]
##   budget_friendly Type   count
##   <lgl>           <chr>  <int>
## 1 FALSE           Cream     15
## 2 FALSE           Liquid     6
## 3 FALSE           Powder    17
## 4 TRUE            Cream      4
## 5 TRUE            Liquid     2
## 6 TRUE            Powder     6
```

Above are the numeric and categorical summary statistics. Review and price were the two numeric variables examined. Among the 50 blushes, the average 5-star review was 4.444 stars, with a standard deviation of about 0.25. This means that many blushes have a relatively high review, just a bit below 5 stars. The median price is $24.50 which can be discouraging if your budget is lower than that. But all is not lost, there are still options available. Assuming the budget is $12, a variable called budget_friendly was added via mutate to see if what blushes were below twelve dollars (TRUE) and which were not (FALSE). Additionally, shoppers have particular preferences for the type of blush they wear. Some prefer cream, others powder, etc. Hence, the mutated variable was grouped with the type of blush (the categorical variable) and the counts were calculated for each combination. There are 4 cream, 2 liquid, 6 powder blushes that are budget friendly.

4. Tidying

Above and below average brands

```
#Showing above avg first
Blush_data %>%
  arrange(desc(Review))%>%
  select(-Name)%>%
  mutate(better_review = Review>4.444)%>%
    pivot_wider(names_from = better_review, values_from = Brand)%>%
  rename(c("Below Avg Review" = "FALSE", "Above Avg Review" = "TRUE"))
```

```
## # A tibble: 50 x 6
##    Price Review Colors Type   `Above Avg Review`      `Below Avg Review`
##    <dbl>  <dbl>  <dbl> <chr>  <chr>                   <chr>
##  1 38       4.8      8 Powder Pat McGrath Labs        <NA>
##  2 16       4.8      6 Cream  LYS Beauty              <NA>
##  3 33       4.8      5 Powder Lancome                 <NA>
##  4 40       4.7     23 Powder Nars                    <NA>
##  5 32       4.7      9 Powder Laura Mercier           <NA>
##  6 32       4.7      5 Cream  Anastasia Beverly Hills <NA>
##  7 30       4.7      6 Cream  Rose Inc                <NA>
##  8 26.5     4.7      5 Powder Clinique                <NA>
##  9 26.5     4.7      8 Powder Clinique                <NA>
## 10  9.99    4.7      4 Cream  Milani                  <NA>
## # ... with 40 more rows
```

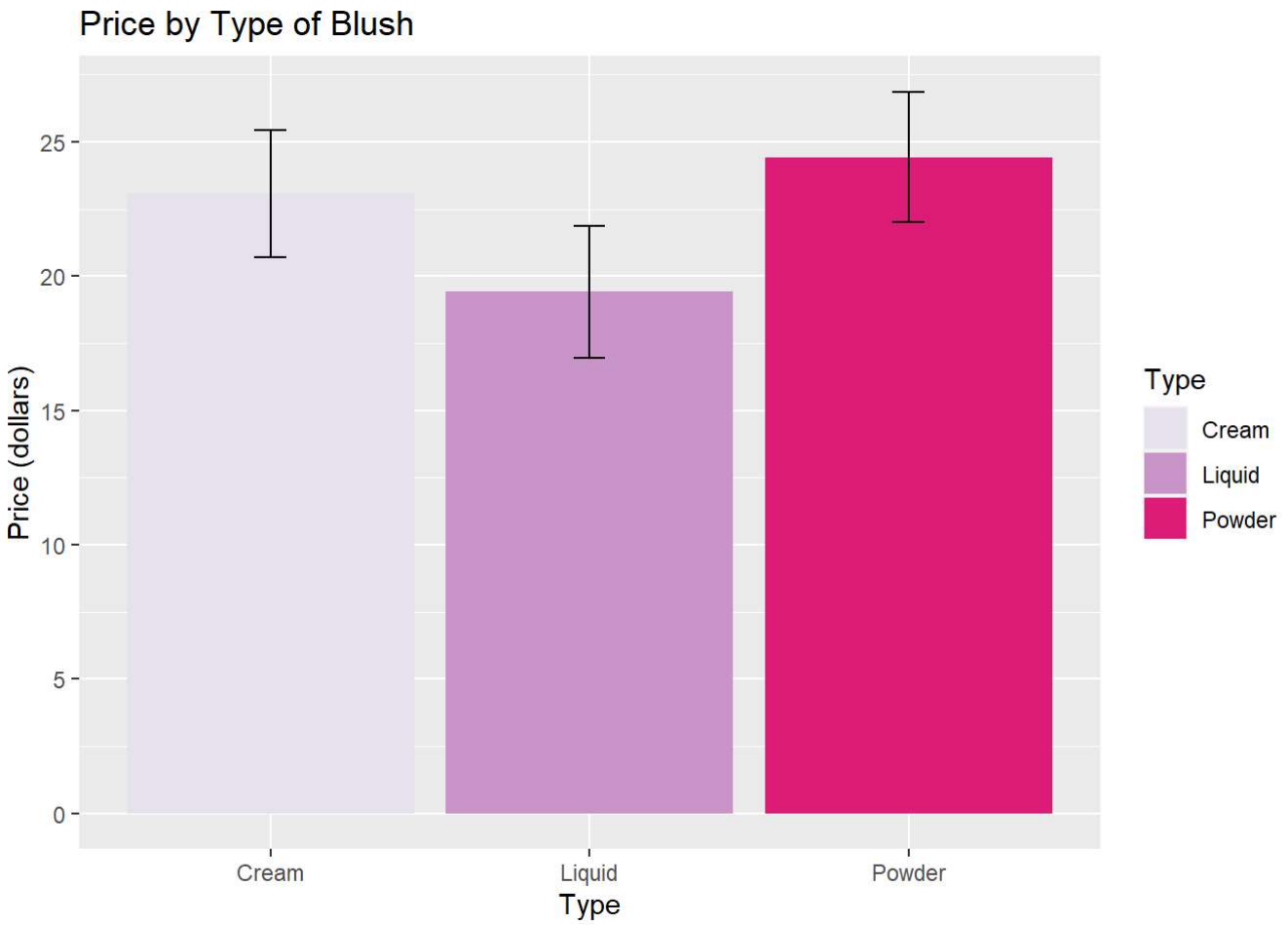
```
##Showing below avg first
Blush_data %>%
  arrange(Review)%>%
  select(-Name)%>%
  mutate(better_review = Review>4.444)%>%
    pivot_wider(names_from = better_review, values_from = Brand)%>%
  rename(c("Below Avg Review" = "FALSE", "Above Avg Review" = "TRUE"))
```

```
## # A tibble: 50 x 6
##    Price Review Colors Type   `Below Avg Review`          `Above Avg Review`
##    <dbl>  <dbl>  <dbl> <chr>  <chr>                       <chr>
##  1  8       3.6      8 Powder ColourPop                   <NA>
##  2  4       3.9      6 Cream  e.l.f. Cosmetics            <NA>
##  3  6.49    3.9      8 Powder Maybelline                  <NA>
##  4 20       4.1      8 Cream  Milk Makeup                 <NA>
##  5  9.99    4.1      4 Powder Milani                      <NA>
##  6 18       4.1      5 Liquid Benefit Cosmetic            <NA>
##  7 30       4.2      3 Liquid Nars                        <NA>
##  8 34       4.2      7 Cream  ILIA                        <NA>
##  9 21       4.3      5 Cream  Rare Beauty by Selena Gomez <NA>
## 10 22       4.3     10 Cream  Fenty Beauty by Rihanna     <NA>
## # ... with 40 more rows
```

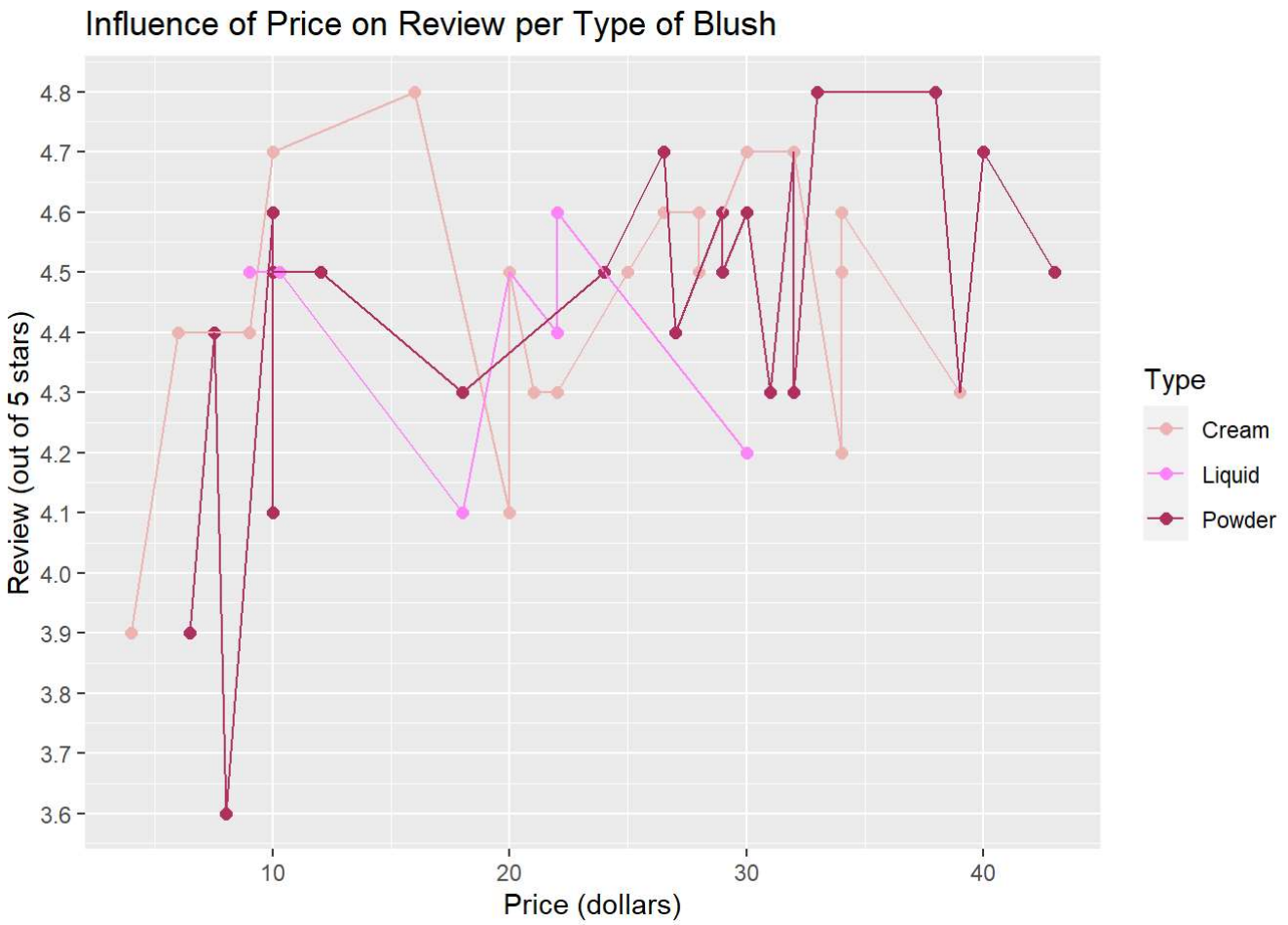
The above table organizes brands into either above or below the average review (which was calculated before as 4.444 stars). The 'Name' variable was taken out as it seemed irrelevant, just the brand was needed. The other variables were left in the table just in case someone wanted to know more about the particular brand's blushes. A new variable was mutated in which states whether or not a review was above ("TRUE") or below ("FALSE") the average. This was pivoted so that 'above' and 'below' average could have their own columns with the brand names underneath for better viewing. However this leaves columns titled "TRUE" and "FALSE", so to better understand what the columns mean, the columns were renamed from "TRUE" to "above avg review" and "FALSE" to "below avg review". This was repeated but with the default ascending order to show both above as well as below average columns in the report. Notably, there are brands like Milani which are affordable and are considered above average and there are brands like ILIA which are expensive and fall below average. So not every above average brand needs to be expensive, maybe quality isn't necessarily an indicator of price.

5. Visualizations

```
library(ggplot2)
Blush_data %>% ggplot(aes(x=Type, y=Price, fill=Type)) +
  geom_bar(stat = "summary", fun = "mean" ) +
  geom_errorbar(stat = "summary", fun.data = "mean_se", width = 0.1) +
    scale_fill_brewer(palette = "PuRd")+
  labs (title = "Price by Type of Blush", x = "Type", y = "Price (dollars)" )+
  scale_y_continuous(breaks = seq(0,30,5))
```




```
BlushPlot1 <- Blush_data %>% ggplot(aes(x = Price, y = Review, color=Type)) + geom_point(size =
2, alpha = 1)+  geom_line() + ggtitle("Influence of Price on Review per Type of Blush")+ labs(y=
"Review (out of 5 stars)", x = "Price (dollars)") + scale_color_manual(values=c("#EEB4B4", "#FF8
3FA", "#B03060"))+ scale_y_continuous(breaks = seq(3.5,5,0.1))
print(BlushPlot1)
```



Above are two graphs. The bar graph shows the mean number of colors by the type of blush, and the scatterplot shows the effect of price on the five-star review, and is color coded by the type of blush.

The bar graph shows the average price for each type of blush, with liquid being the cheapest and powder as the most expensive. However, the overlapping error bars indicate that the differences in the mean price between the different types of blushes are insignificant. But this good news. There is no trend between the type of blush and its price. This shows that while cream blushes have been booming in recent years, the production in both high-end as well as drugstore options fortunately allows us to shop more trendy makeup without breaking the bank. This also means that regardless of your preference for a type of blush to shop for, be it cream, liquid, or powder, you aren't necessarily missing out on a shopping deal.

The scatterplot displays, for the most part, insignificant data. There does not appear to be a relationship between the price and the five-star review, and there is no trend based on product type either. While the graph, does seem to have higher reviews, the more expensive the product, there is too much variation to conclude anything as significant. Interestingly, the cheapest products (less than ten dollars), have noticeable decrease in review rating. This could perhaps indicate that once a blush product is worth less than ten dollars there may be a decrease in quality. However, a big caveat to this includes a confounding variable: shoppers psychologically could be convincing themselves that the more expensive a product is, the better its quality, so that conclusion cannot be supported.