GDSC - Fashion Trend Analysis Project Report

2023-05-31

To investigate the relationship between trends within different sectors of fashion, and see how they influence one another, we collected data and performed some experiments to test some of our hypotheses. We started by manually collecting data from a variety of sources, and creating CSV files that could be used in our report. We acknowledge that manual data collection involves some human error, but this step was crucial for learning the entire process of conducting experiments, and for coming up with appropriate hypotheses. In the future, this data collection could be automated using web-scraping, pattern recognition, and an AI/ML model, which is something our team plans to pursue further in the future.

Getting Started

The following are the necessary packages we use within our testing:

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(tinytex)
library(rcompanion)
```

Next, we create data frames from each CSV file containing separately collected data, allowing for easy implementation and analysis:

```
depop <- data.frame(read.csv("Depop Data Collection - Sheet1.csv"))

hm <- data.frame(read.csv("Hannah Wen Data Final - Sheet1.csv"))

shein <- data.frame(read.csv("Hannah Wen Data Final - Sheet2.csv"))

highfashion_dresses <- data.frame(read.csv("High Fashion Data Collection - dress trends.csv"))

highfashion_spring <- data.frame(read.csv("High Fashion Data Collection - overall spring trends.csv"))

pinterest <- data.frame(read.csv("Pinterest Data Collection - Sheet1.csv"))
```

```
nordstrom <- data.frame(read.csv("Zara_Nordstrom Dataset - nordstrom (1).csv"))

zara <- data.frame(read.csv("Zara_Nordstrom Dataset - zara (1).csv"))
```

First Experiment:

Are midi dresses the most on-trend amongst all clothing retailers for current spring/summer season?

Hypothesis: Midi dresses are more common than other dress lengths for all retailers.

In order to test this hypothesis we will run a Goodness of Fit test for each retailer.

First we have to ensure that all requirements are met for the test.

Chi Squared Goodness of Fit Test Requirements:

- Samples are independent: there is human error because data was collected manually but they are all randomly selected from the "Trending" or recommended pages of the retailer
- The data is categorical with the quality we are interested in being dress length with levels: mini, midi, maxi -Expected values are greater than 5 for each category. We have three categories so 3*5=15 so n must be >= 15

Formatting sets in a way that makes them easier to manipulate later:

```
Retailer = c("Zara")
zara2 <- zara %>%
  rename("Length" = "Length..dress.skirts.")
zaraF <- cbind(zara2, Retailer)%>%
    filter(Article == "dress")

nordstrom2 <- nordstrom %>%
  rename("Length" = "Length..dress.skirts.")
Retailer=c("Nordstrom")
nordstromF <- cbind(nordstrom2, Retailer)%>%
    filter(Article == "dress")

hmF <- hm %>%
  rename("Length" = "Length..dress.skirts.")%>%
  filter(Article == "dress")

sheinF <- shein %>%
  rename("Length" = "Length..dress.skirts.")

Retailer = c("Depop")
depopF <- cbind(depop, Retailer)%>%
    filter(Article == "dress")

Retailer = c("Pinterest")
pinterestF <- cbind(pinterest, Retailer)%>%
  filter(Article == "dress")

highF <- highfashion_dresses %>%
  filter(article == "dress")
```

Creating frequency tables for mini, midi, and maxi length dresses to see if the necessary sample size is met for each data set:

```
zaraCount <- zaraF %>%
  group_by(Length)%>%
  summarize(count = n())
sum(zaraCount$count)
```

```
## [1] 17
```

```
nordstromCount <- nordstromF%>%
  group_by(Length)%>%
  summarize(count = n())
sum(nordstromCount$count)
```

```
## [1] 8
```

```
depopCount <- depopF %>%
  group_by(Length) %>%
  summarize(count=n())
sum(depopCount$count)
```

```
## [1] 16
```

```
pinCount <- pinterestF %>%
  group_by(Length)%>%
  summarize(count = n())
sum(pinCount$count)
```

```
## [1] 3
```

```
sheinCount <- sheinF%>%
  group_by(Length) %>%
  summarize(count=n())
sum(sheinCount$count)
```

```
## [1] 25
```

```
hmCount <- hmF%>%
  group_by(Length) %>%
  summarize(count=n())
sum(hmCount$count)
```

```
## [1] 21
```

```
highCount<- highfashion_dresses%>%
  group_by(length) %>%
  summarize(count=n())
sum(highCount$count)
```

```
## [1] 30
```

Zara: n = 17 Nordstrom: n = 8 Depop: n = 16 Pinterest: n = 3 SHEIN: n=11 H&M: n = 21 High Fashion: n = 30

We find that the third requirement of expected value is met for Zara, Depop, H&M, and high fashion. Unfortunately, the remainder of the retailers/data do not meet the requirements, and will therefore be left out of this specific hypothesis testing. Since our data was not collected to accomodate this number of dresses, a potential point of growth could be collecting more dress data from those brands.

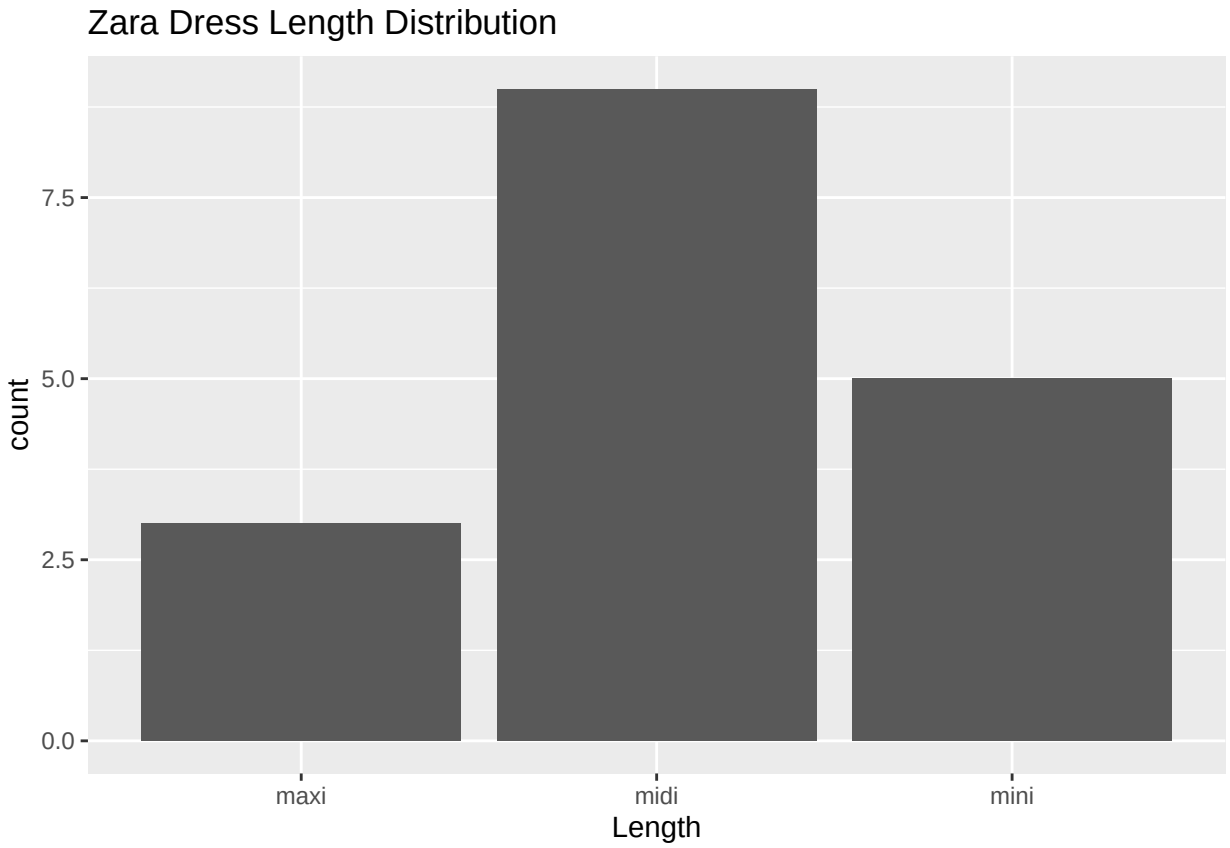
H_0: The proportion of mini, midi, and maxi length dresses are the same. H_A: The proportion of mini, midi, and maxi length dresses are not the same.

α=0.05 (significance level)

The following will plot the distribution of lengths, as well as perform a chi-squared test for each source of data (retailer).

Zara Dresses:

```
g <- ggplot(zaraF, aes(Length))
g+geom_bar()+ggtitle("Zara Dress Length Distribution")
```




```
resZara <-chisq.test(zaraCount$count, p=rep(1/length(zaraCount$Length), 3))
resZara
```

```
##
```

```
##  Chi-squared test for given probabilities
##
## data:  zaraCount$count
## X-squared = 3.2941, df = 2, p-value = 0.1926
```

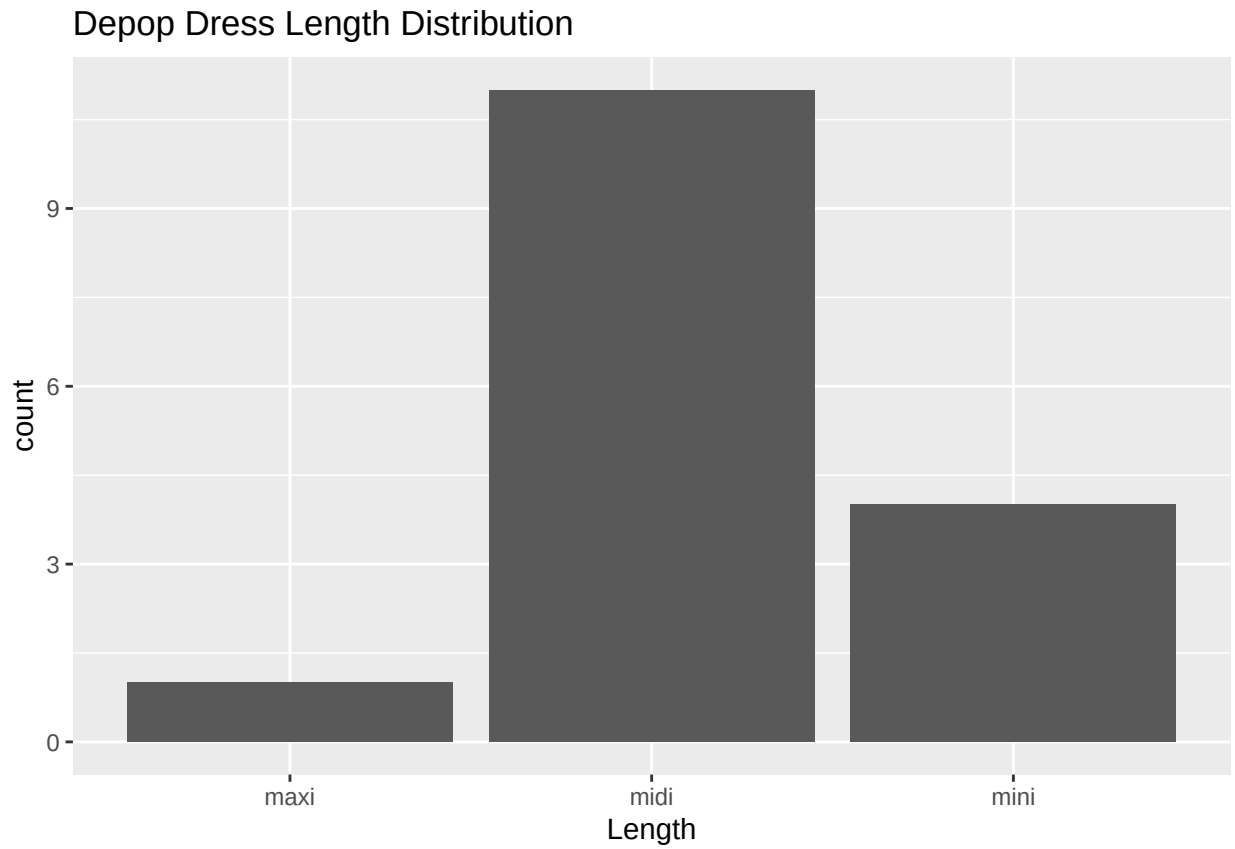
p-value $> \alpha$

Since the p-value is greater than the alpha level we fail to reject the null hypothesis and conclude that the proportions of mini, midi, and maxi length dresses are the same. This means that the variation in the frequency of dress length in our sample size does not necessarily mean that there is a variation in frequency in the population.

This disproves our original question of whether midi dresses are the most popular dress length amongst all retailers, however, now we can analyze the trends of individual retailers.

Depop Dresses:

```
g <- ggplot(depopF, aes(Length))
g+geom_bar() + ggtitle("Depop Dress Length Distribution")
```




```
resDepop <-chisq.test(depopCount$count, p=rep(1/length(depopCount$Length), 3))

resDepop
```

```
##
##  Chi-squared test for given probabilities
```

```
##
## data:  depopCount$count
## X-squared = 9.875, df = 2, p-value = 0.007173
```

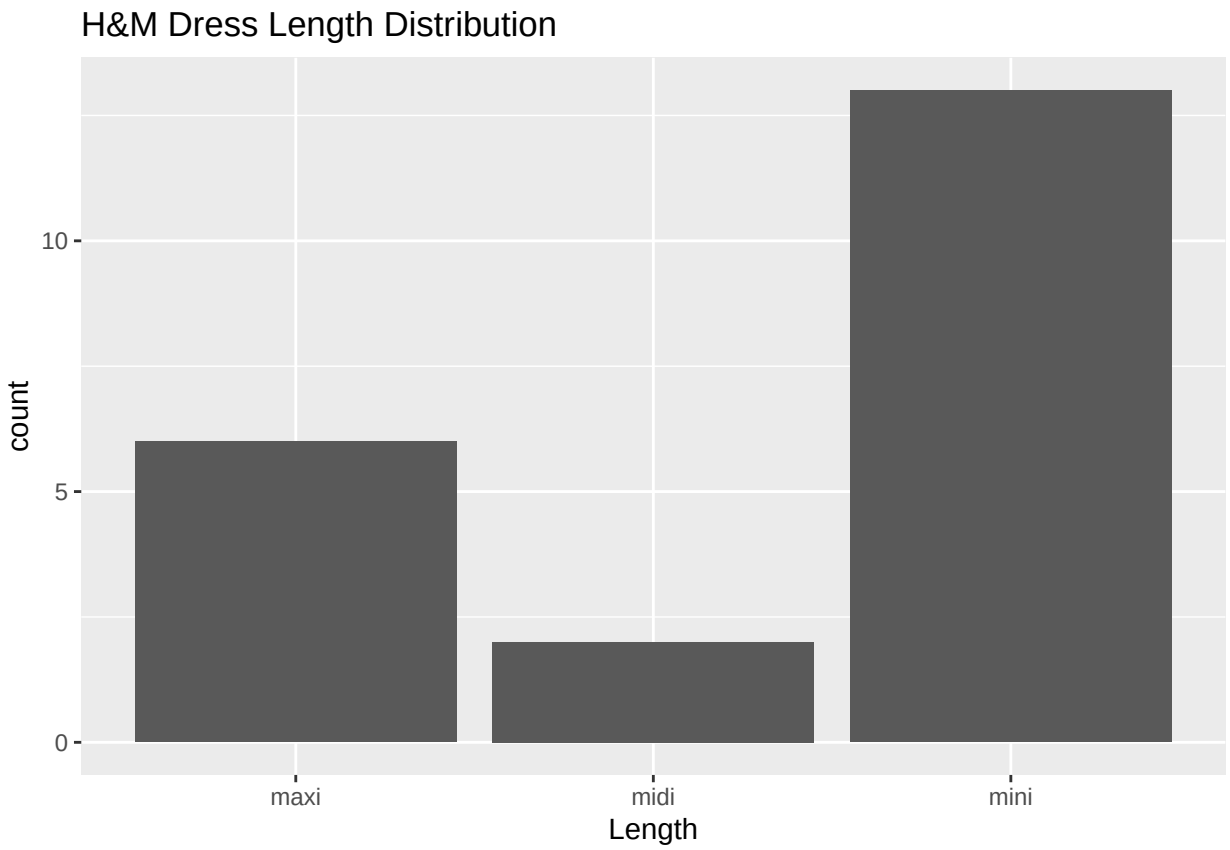
p-value $< \alpha$

Since the p-value is less than the alpha level we reject the null hypothesis and conclude that the proportion of mini, midi, and maxi length dresses are not the same.

We can see in the histogram that there is a larger number of midi-length dresses than there are mini and maxi which can lead us to infer that that length is the most popular dress length on Depop in comparison to the other lengths.

H&M dresses:

```
g <- ggplot(hmF, aes(Length))
g+geom_bar()+ggtitle("H&M Dress Length Distribution")
```




```
reshm <-chisq.test(hmCount$count, p=rep(1/length(hmCount$Length), 3))
reshm
```

```
##
##  Chi-squared test for given probabilities
##
## data:  hmCount$count
## X-squared = 8.8571, df = 2, p-value = 0.01193
```

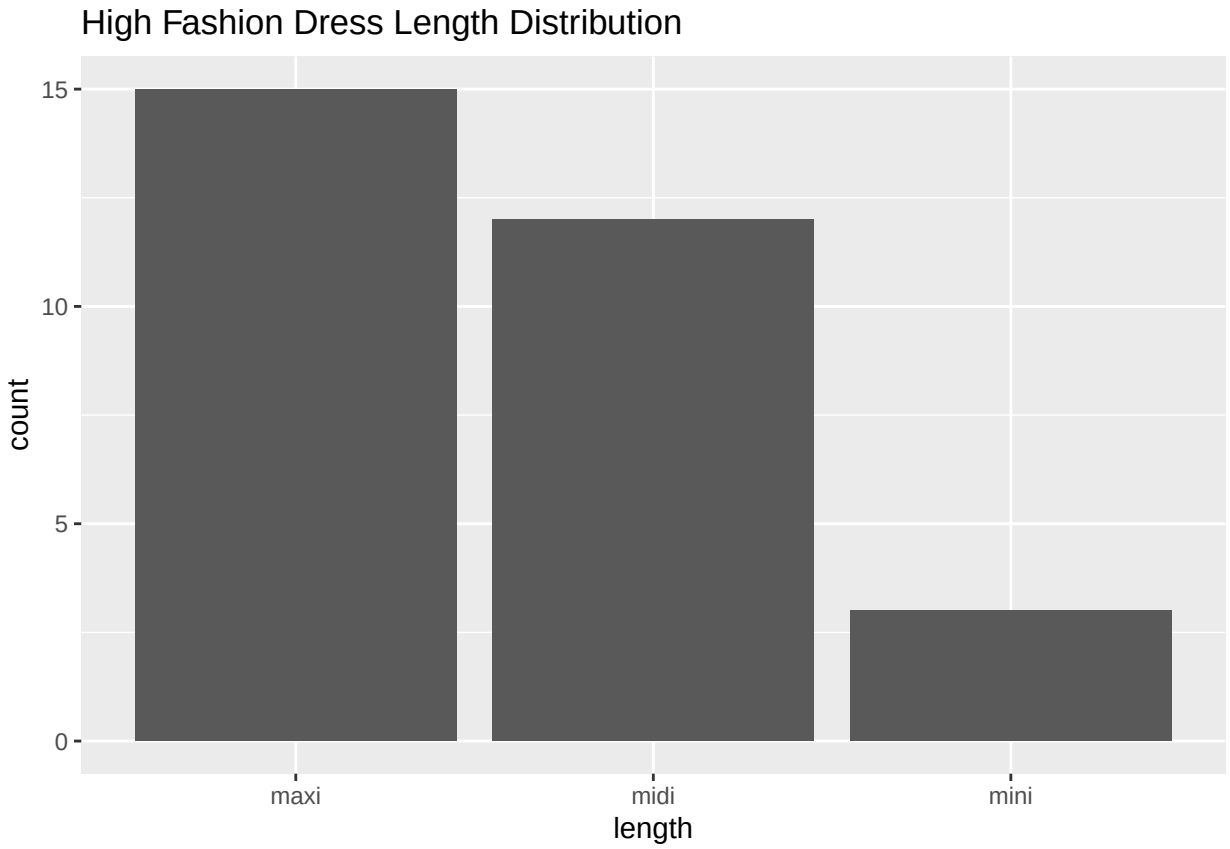
p-value $< \alpha$

Since the p-value is less than the alpha level we reject the null hypothesis and conlude that the proportion of mini, midi, and maxi length dresses are not the same.

Here we see that there is a larger number of mini dresses so we can infer that this is the variable that is causing us to reject the null hypothesis. However, this disproves our original hypothesis that midi length dresses are the most popular among all retailers as this test supports that the most popular dress length at H&M is mini.

High Fashion Dresses:

```
g <- ggplot(highF, aes(length))
g+geom_bar()+ggtitle("High Fashion Dress Length Distribution")
```




```
resHigh <-chisq.test(highCount$count, p=rep(1/length(highCount$length), 3))
resHigh
```

```
##
##  Chi-squared test for given probabilities
##
## data:  highCount$count
## X-squared = 7.8, df = 2, p-value = 0.02024
```

p-value $< \alpha$

Since the p-value is less than the alpha level we reject the null hypothesis and conclude that the proportion of mini, midi, and maxi length dresses are not the same.

Both maxi and midi dresses have relatively high frequencies in this distribution so we can infer that those are the variables causing us to reject the null hypothesis. However, they are relatively even in number so it is inconclusive whether midi or maxi length dresses is most popular although we can conculsively infer that mini dresses are the least popular.

Conclusion About Dress Length:

Despite having noticed a lot of midi dresses in the process of our data collection, our results showed that the only retailer where midi dresses are most on-trend is Depop. This is an interesting observation that can give us information on how trends are evolving. Depop is a second-hand retailer that sells primarily to young-adults; this may indicate that midi dresses are being produced less by modern fashion retailers, but are potentialy coming back in-style for young adults, as they are being sold more used than new.

Second Experiment:

At what price-point can we find the most skirts and dresses with pockets? What does this say about the cost of comfort in women's clothing?

We looked at different clothing retailers at various prices, ranging from $5 pieces at SHEIN and $5000 pieces from high fashion brands, to see where we are most likely to find women's skirts and dresses with pockets.

```
# no dresses or skirts with pockets
depop %>% filter(Article %in% c('dress', 'skirt'))
```

```
##    Article       Fit Material        Color Embellishment Sleeve.type Neckline
## 1    dress  straight  crochet multicolored       pattern  sleeveless    round
## 2    dress    fitted    rayon         pink        floral  sleeveless    round
## 3    dress     loose     <NA> multicolored       gingham       short collared
## 4    dress   fitted      <NA>        black       pattern  sleeveless   v-neck
## 5    dress     loose     <NA>        black          <NA>  sleeveless    heart
## 6    dress     loose     <NA>        green        floral  sleeveless   v-neck
## 7    dress     loose     <NA>         pink          bows       short    round
## 8    dress  straight   velvet        black       pattern        long    round
## 9    dress    fitted   cotton        white          <NA>  sleeveless   v-neck
## 10   dress    fitted     <NA>        black         sheer        long   v-neck
## 11   dress     loose     <NA>          red         sheer  sleeveless    round
## 12   dress    fitted     <NA>        green          <NA>       short   v-neck
## 13   dress    fitted     <NA>        black        floral       strap    cowel
## 14   dress    fitted     <NA>        green          <NA>       short  square
## 15   dress fit-flare     <NA>         grey        layers      halter straight
## 16   dress     loose     <NA>        white        floral      straps  square
##    Cropped Length Rise Pockets
## 1       NA   mini <NA>   FALSE
## 2       NA   midi <NA>   FALSE
## 3       NA   midi <NA>   FALSE
## 4       NA   midi <NA>   FALSE
## 5       NA   midi <NA>   FALSE
## 6       NA   midi <NA>   FALSE
## 7       NA   midi <NA>   FALSE
```

```
## 8       NA   midi <NA>   FALSE
## 9       NA   mini <NA>   FALSE
## 10      NA   midi <NA>   FALSE
## 11      NA   mini <NA>   FALSE
## 12      NA   midi <NA>   FALSE
## 13      NA   maxi <NA>   FALSE
## 14      NA   midi <NA>   FALSE
## 15      NA   mini <NA>   FALSE
## 16      NA   midi <NA>   FALSE
```

```
# no dresses or skirts with pockets among h&m or shein

highfashion_dresses %>% count(pockets) # 3 dresses with pockets out of 30
```

```
##   pockets  n
## 1       n 27
## 2       y  3
```

```
highfashion_spring %>% filter(article %in% c('dress', 'skirt')) %>% count(pockets) # 2/11
```

```
##   pockets n
## 1       n 9
## 2       y 2
```

```
pinterest %>% filter(Article %in% c('dress', 'skirt')) # no pockets among these in pinterest
```

```
##   Article   Fit Material  Rise Embellishment Length     Sleeve Neckline Pockets
## 1   dress tight                               mini  sleveless    scoop
## 2   skirt tight           high     split hem                                no
## 3   dress tight                               mini sleeveless straight
## 4   dress loose          hight                midi              floral      no
##   Color               Search
## 1 black        trendy outfits
## 2 white trending outfits 2023
## 3 cream        spring outfits
## 4  pink        spring outfits
```

```
nordstrom %>% filter(Article %in% c('dress', 'skirt')) # no pockets
```

```
##   Article     Fit  Material Rise..pants.skirts.  Embellishment
## 1   dress  fitted      <NA>               <NA>           <NA>
## 2   dress  fitted      <NA>               <NA>           <NA>
## 3   dress relaxed      <NA>               <NA>         A-line
## 4   dress  fitted      <NA>               <NA>           <NA>
## 5   dress  fitted      <NA>               <NA>           <NA>
## 6   dress relaxed polyester               <NA>        ruffles
## 7   dress  fitted      <NA>               <NA> sheath/ruching
## 8   dress   loose     rayon               <NA>           <NA>
##   Length..dress.skirts.     Sleeve.type         Neckline Pockets
## 1                 maxi spaghetty strap           scoop     <NA>
## 2                 maxi            wide           v-neck    <NA>
```

```
## 3                  midi     puff sleeve off the shoulder    <NA>
## 4                  midi      one-sleeve     asymmetrical    <NA>
## 5                  midi  flutter sleeve           v-neck    <NA>
## 6                  midi         flutter            jewel    <NA>
## 7                  midi      sleeveless       scoop neck    <NA>
## 8                  maxi  spaghetti stra           v-neck    <NA>
##                                 Color
## 1           gray/red/brown/nude/olive
## 2           red/blue/black/green/pink
## 3                        orange/multi
## 4    orange/black/blue/pink/white/red
## 5                      black/red/blue
## 6         black/blue/green/pink/white
## 7 pink/green/blue/black/white/red/tan
## 8    black/green/blue/red/pink/yellow
##
## 1   https://www.nordstrom.com/s/skims-soft-lounge-long-slipdress-regular-plus-size/5897391?origin=ca
## 2             https://www.nordstrom.com/s/v-neck-jersey-maxi-dress-regular-petite/3965546?origin=ca
## 3                       https://www.nordstrom.com/s/off-the-shoulder-a-line-dress/7312428?origin=ca
## 4 https://www.nordstrom.com/s/dress-the-population-tiffany-one-shoulder-midi-dress/5064746?origin=ca
## 5     https://www.nordstrom.com/s/maggy-london-flutter-sleeve-faux-wrap-midi-dress/5604756?origin=ca
## 6             https://www.nordstrom.com/s/cece-clip-dot-flutter-sleeve-midi-dress/7045421?origin=ca
## 7       https://www.nordstrom.com/s/ruched-side-sleeveless-dress-regular-plus-size/5795330?origin=ca
## 8                                https://www.nordstrom.com/s/woven-favorite-dress/6267014?origin=ca
```

```
zara %>% filter(Article %in% c('dress', 'skirt')) %>% count(Pockets) # 2 out of 19
```

```
##   Pockets  n
## 1     Yes  2
## 2    <NA> 17
```

Data observations pulled from Depop, Shein, H&M, Nordstrom, and Pinterest did not have any dresses or skirts with pockets. Dresses pulled from the high fashion dresses data frame demonstrated a 10% pocket rate with 3 out of 30 dresses having pockets. Skirts and dresses pulled from the general high fashion spring trends data frame demonstrated a 18% pocket rate with 2 out of 11 articles having pockets. Zara has approximately a 10% pocket rate with 2 articles out of 19 having pockets.

This implies that pockets in skirts and dresses are more common at a higher price level, while some of the cheapest retailers had zero dresses or skirts with pockets. However, it is important to observe that the sample size in each data frame is small, and that the high fashion data frames pull from different brands while data frames like the Zara or H&M one pull from only one. In general, it is observed that pockets are a rare occurrence among skirts and dresses.

Pockets Conclusion:

Seeing as that pockets are more likely to be found in high-end fashion, they can somewhat be considered a "luxury," and many women would probably agree. The proportions were all very low, showing that it is extremely rare to find dresses or skirts with pockets. Although we only investigated women's clothing, it would be very interesting to see how common pockets are in men's clothing in comparison. This future experiment could reveal a lot about the lack of priority that comfort has in women's fashion compared to men's fashion, and what this means in modern society.

Third Experiment:

How similar are the materials that are used among different retailers at different price points?

Most retailers sell clothes that consist of synthetic fibers such as polyester, viscose, and rayon. Garments made of these materials will not be considered since they are so popular, especially among fast fashion retailers. While some data frames such as the high fashion data frame recorded every single material, some did not mention the extremely common materials. In the SHEIN data frame, no material was recorded since every garment observed was some combination of the common materials (polyester, viscose, rayon). Thus, it is excluded from these tests. A chi-squared test of independence is typically performed when testing whether two categorical variables are independent. However, do to the small sample size, we will use the Fisher's exact probability test.

We need to create proportions for materials, and set up the hypothesis test.

Data Manipulation:

```
# pulling vectors of materials for each category
depop_materials = depop %>% pull(Material)
hm_materials = hm %>% pull(Material)
# no materials for shein recorded because there was no material observed that was not one of the typica
highfashion_dress_materials = highfashion_dresses %>% pull(material)
highfashion_spring_materials = highfashion_spring %>% pull(material)
pinterest_materials = pinterest %>% pull(Material)
nordstrom_materials = nordstrom %>% pull(Material)
zara_materials = zara %>% pull(Material)
```

```
table(depop_materials)
```

```
## depop_materials
##    cargo   corset   cotton  crochet    denim     lace    linen     mesh
##        1        2        6        3        6        1        1        2
##   mohair    Nylon    rayon    satin tapestry   velvet
##        1        1        1        1        1        1
```

```
table(hm_materials)
```

```
## hm_materials
## chiffon  cotton  crepe  crochet  jersey   satin
##       1       2       2       2       3       2
```

```
table(highfashion_dress_materials)
```

```
## highfashion_dress_materials
##       chiffon        cotton cotton; linen        jersey         linen
##             1             6             1             1             4
##     polyester        poplin         satin          silk  silk; cotton
##             1             8             1             2             1
##       taffeta       viscose viscose; silk
##             1             1             2
```

```
table(highfashion_spring_materials)
```

```
## highfashion_spring_materials
##       bamboo lyocell               cotton                denim
##                    1                    5                    3
##                gauze               jersey                nylon
##                    1                    6                    2
##               poplin                 silk               twill
##                    1                    1                    1
##              viscose viscose; rayon; wool        viscose; wool
##                    4                    1                    1
##                 wool
##                    2
```

```
table(pinterest_materials)
```

```
## pinterest_materials
##        cotton  denim  satin
##     28      3      2      1
```

```
table(nordstrom_materials)
```

```
## nordstrom_materials
##      chiffon       cotton        denim        linen linen/cotton    polyester
##            1            5            1            5            1            4
##        rayon
##            1
```

```
table(zara_materials)
```

```
## zara_materials
##      cotton       linen linen blend   polyester       tulle
##          15           4           3           4           5
```

Now that we can find all the counts for each material we want to perform a test on, we can begin testing. Taking a general look of the contents of the materials, we will calculate the proportions of the materials: silk/satin, cotton/linen, and jersey.

Due to the small sample size, Fisher's Exact Test will be performed. This test requires the assumptions that the data collected is a random sample, that the data points are dependent of each other, and that the groups of our categorical variable are mutually exclusive, meaning that a material cannot be both found and not found in a garment.

The Fisher Exact Test will return a p-value, which is the probability that the null hypothesis is true. Our null hypothesis for each test will be that the proportion of a certain material is constant among each type of retailer. This test will be performed at the commonly used 0.05 level of significance, meaning that if the p-value is less than 0.05, there is sufficient evidence to reject the null hypothesis. The alternative hypothesis is that the proportions are not the same among every retailer.

Calculations and performing the Fisher test:

```
silk.satin.n = c(1, 2, 5, 1, 1, 0, 0) # count of silk or satin found in each dataframe
n = c(length(depop_materials), length(hm_materials), length(highfashion_dress_materials),
      length(highfashion_spring_materials), length(pinterest_materials), length(nordstrom_materials),
      length(zara_materials)) # vector of length of each of the material vectors
names = c("Depop", 'H&M', 'High fashion dresses', 'High fashion spring trends',
               'Pinterest', 'Nordstrom', 'Zara') # row names
silk.satin = data.frame("Silk/Satin" = silk.satin.n, "Silk/satin not found" = n - silk.satin.n,
                        row.names = names)
silk.satin
```

```
##                             Silk.Satin Silk.satin.not.found
## Depop                                1                   48
## H&M                                  2                   23
## High fashion dresses                 5                   25
## High fashion spring trends           1                   28
## Pinterest                            1                   33
## Nordstrom                            0                   39
## Zara                                 0                   49
```

```
fisher.test(silk.satin)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  silk.satin
## p-value = 0.005258
## alternative hypothesis: two.sided
```

The p-value is approximately 0.005, which means that we reject the null hypothesis that the proportion of satin and/or silk found in garments is not the same among all retailers.

```
cotton.linen.n = c(6, 2, 12, 5, 3, 11, 22) # vector of count of cotton or linen found
cotton.linen = data.frame("Cotton/Linen" = cotton.linen.n, "Cotton/linen not found" = n - cotton.linen.n,
                row.names = names)
cotton.linen
```

```
##                             Cotton.Linen Cotton.linen.not.found
## Depop                                  6                     43
## H&M                                    2                     23
## High fashion dresses                  12                     18
## High fashion spring trends             5                     24
## Pinterest                              3                     31
## Nordstrom                             11                     28
## Zara                                  22                     27
```

```
fisher.test(cotton.linen, workspace=2e8)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  cotton.linen
## p-value = 7.933e-05
## alternative hypothesis: two.sided
```

The p-value is extremely close to 0, providing sufficient evidence to reject the null hypothesis and suggest that the proportions of cotton and/or linen are not the same among each retailer.

```
jersey.n = c(0, 3, 1, 6, 0, 0, 0) # vector of counts of jersey
jersey = data.frame("Jersey" = jersey.n, "Jersey not found" = n - jersey.n,
                row.names = names)
jersey
```

```
##                             Jersey Jersey.not.found
## Depop                            0               49
## H&M                              3               22
## High fashion dresses             1               29
## High fashion spring trends       6               23
## Pinterest                        0               34
## Nordstrom                        0               39
## Zara                             0               49
```

```
fisher.test(jersey)
```

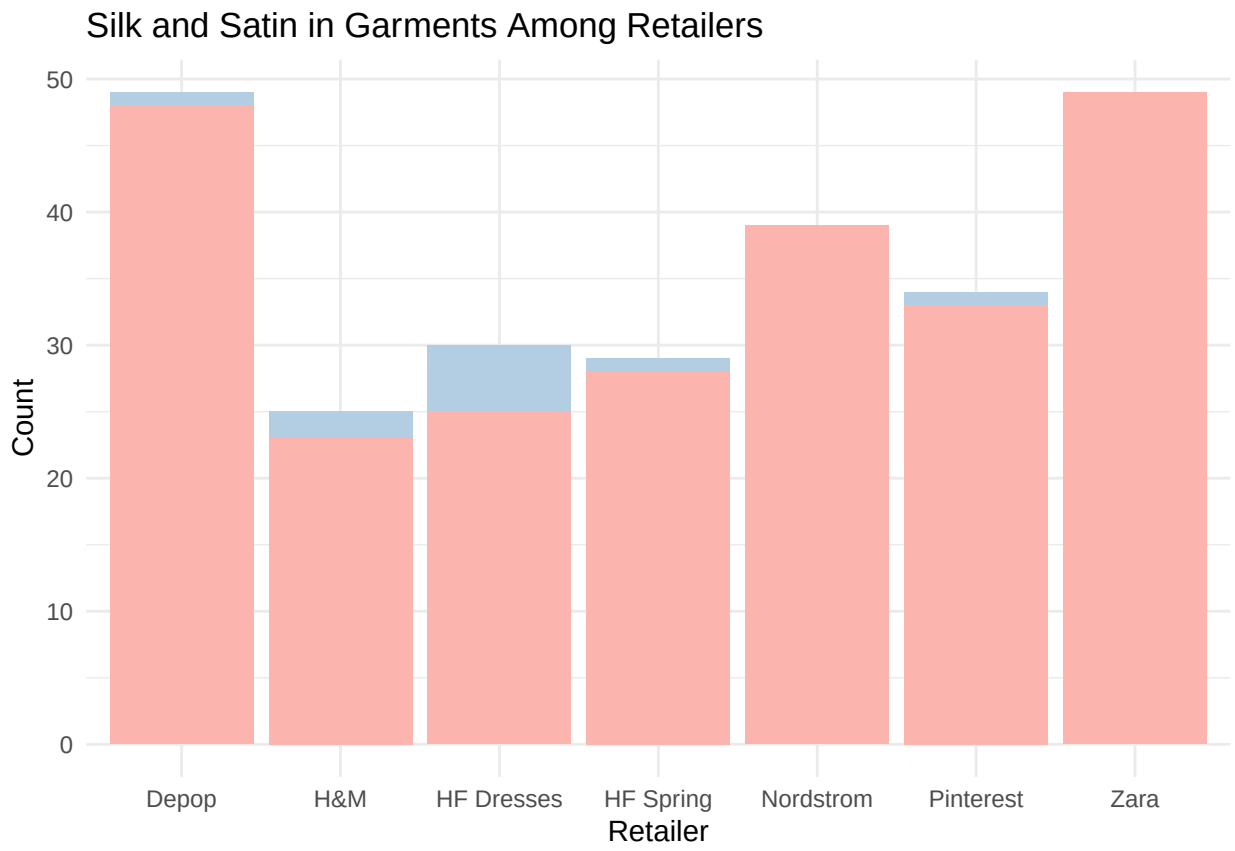
```
##
##  Fisher's Exact Test for Count Data
##
## data:  jersey
## p-value = 1.048e-05
## alternative hypothesis: two.sided
```

The p-value is incredibly close to 0, indicating that we reject the null hypothesis also.

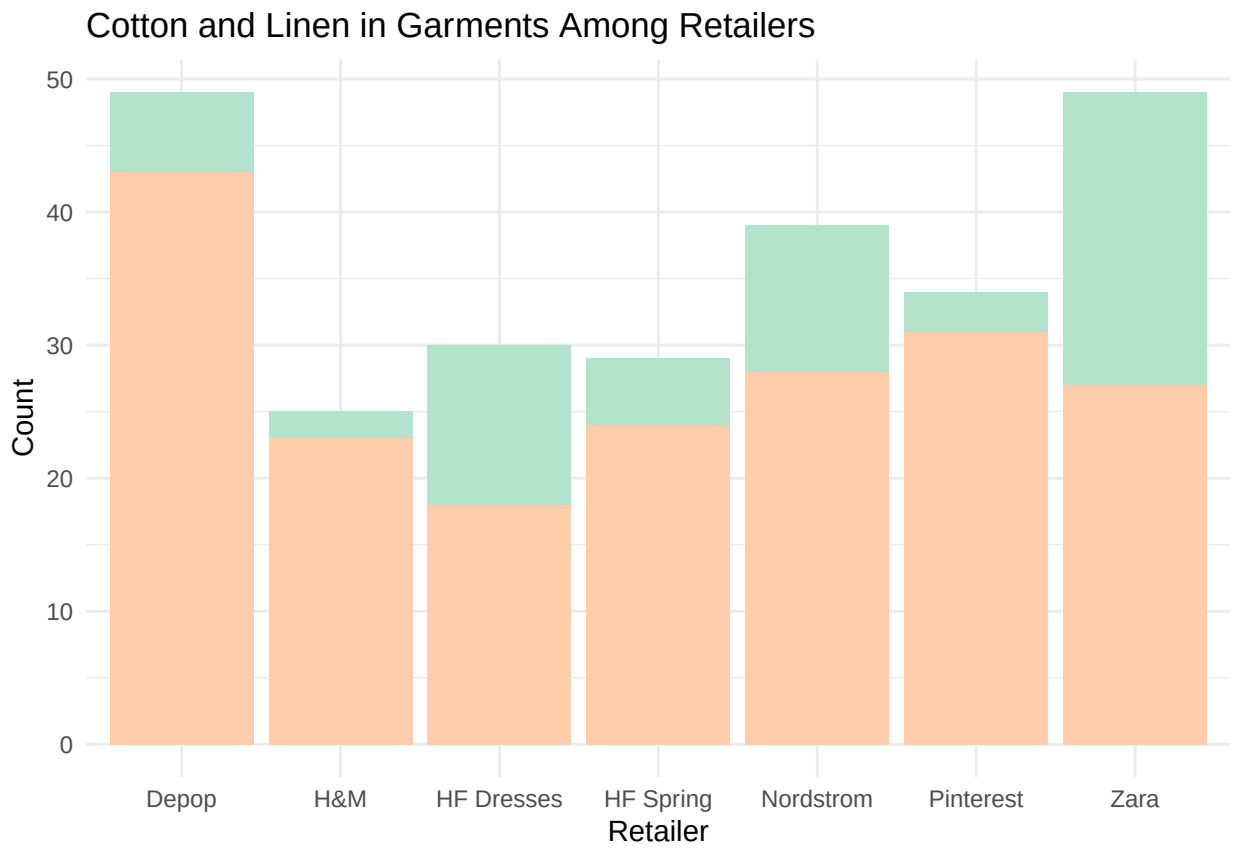
All 3 tests provide sufficient evidence to reject the null hypotheses, suggesting that none of these materials are found at the same rate among every retailer. However, the p-values of jersey and cotton/linen were much smaller than the p-value for satin/silk, meaning that there is a higher chance of the same proportions of satin/silk being found in these retailers again.

Visualizations to Display Material Trends:

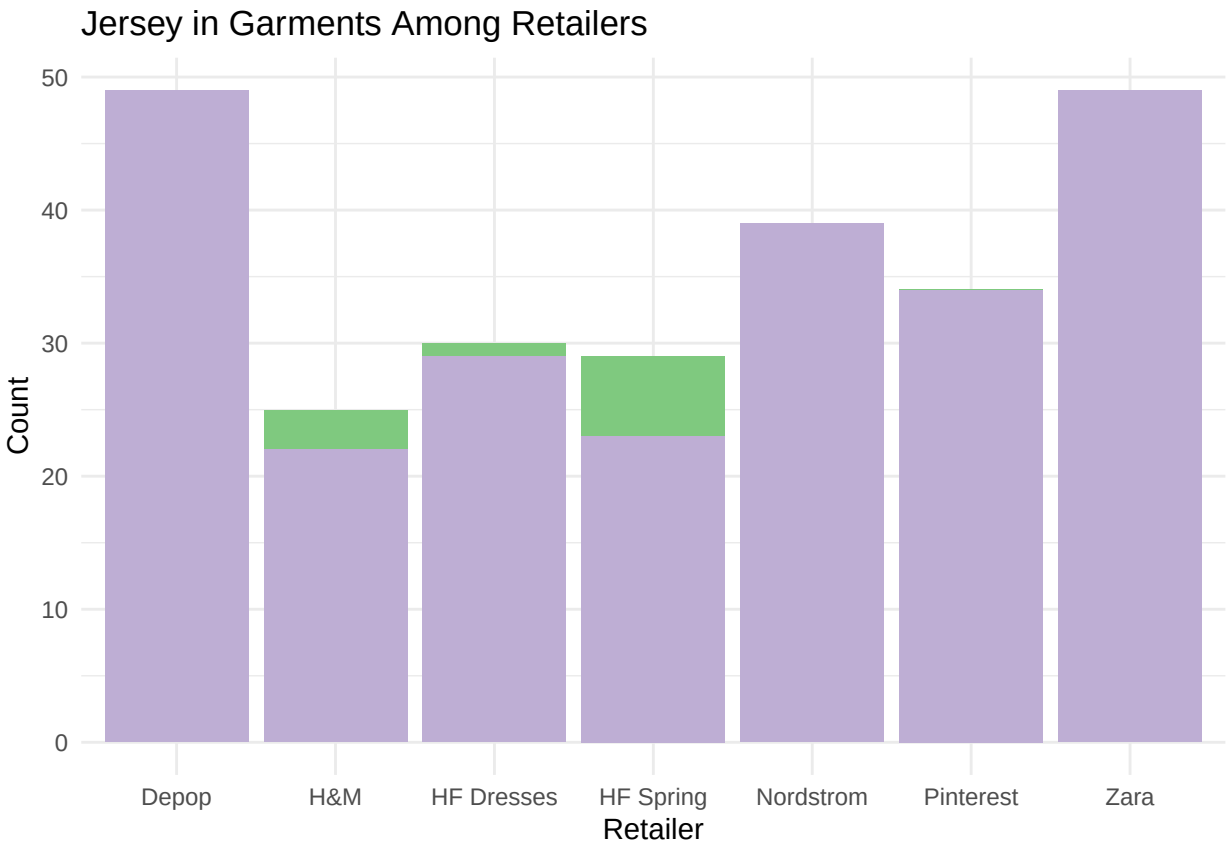
```
X = names = c("Depop", 'H&M', 'HF Dresses', 'HF Spring',
              'Pinterest', 'Nordstrom', 'Zara')
silksat.df = data.frame(X, silk.satin.n, n - silk.satin.n)
silk.sat.df =
  pivot_longer(silksat.df, cols = !X, names_to = 'Type', values_to = 'Count')
ggplot(silk.sat.df, aes(x = X, y = Count, fill = Type)) + geom_bar(stat = 'identity',
        position = position_stack(reverse=TRUE)) + theme_minimal() +
  scale_fill_brewer(palette = 'Pastel1',labels = c('Not found', 'Found')) +
  labs(x = 'Retailer', title = 'Silk and Satin in Garments Among Retailers') +
  theme(legend.position = "none")
```




```
cotton.linen.df = data.frame(X, cotton.linen.n, n - cotton.linen.n)
cotton.linen.df =
  pivot_longer(cotton.linen.df, cols = !X, names_to = 'Type', values_to = 'Count')
ggplot(cotton.linen.df, aes(x = X, y = Count, fill = Type)) + geom_bar(stat = 'identity',
        position = position_stack(reverse=FALSE)) + theme_minimal() +
  scale_fill_brewer(palette = 'Pastel2',labels = c('Not found', 'Found')) +
  labs(x = 'Retailer', title = 'Cotton and Linen in Garments Among Retailers') +
  theme(legend.position = "none")
```




```
jersey.df = data.frame(X, jersey.n, n - jersey.n)
jersey.df =
  pivot_longer(jersey.df, cols = !X, names_to = 'Type', values_to = 'Count')
ggplot(jersey.df, aes(x = X, y = Count, fill = Type)) + geom_bar(stat = 'identity',
        position = position_stack(reverse=FALSE)) + theme_minimal() +
  scale_fill_brewer(palette = 'Accent',labels = c('Not found', 'Found')) +
  labs(x = 'Retailer', title = 'Jersey in Garments Among Retailers') +
  theme(legend.position = "none")
```



Conclusion About Common Materials:

We can see from our testing that certain materials are trending within this spring/summer fashion season, which makes sense in accordance with the weather. Additionally, analyzing proportions of materials seen in different retailers at different price-points may also indicate the higher use of expensive materials in high-fashion compared to cheaper materials in mainstream consumer fashion.

Fourth Experiment:

Is there an association between colors found in high fashion products and colors found in second-hand retail products? If so, to what extent?

Hypothesis: There is a strong positive correlation between colors found in high fashion products and colors found in second-hand retail products. In other words, colors that are seen commonly in high fashion products will also be seen more often in second-hand retail products.

Process:

Conduct a Fisher's Exact Test to determine association between colors found in High Fashion products and colors found in second-hand retail products. If there is an association, run Cramer's V test to measure the association. (Note: I initially attempted to use a Chi-Squared test, but the values per cell were too low. Many values were zero or <=5. For smaller sample sizes, Fisher's test is more suitable.)

Manipulating Data Frames:

Combining the high fashion data frames to create a single one:

```
highfashion <- full_join(highfashion_dresses, highfashion_spring)
```

```
## Joining, by = c("article", "fit", "material", "embellishment", "length",
## "sleeve", "pockets", "brand", "color")
```

```
highfashion <- highfashion %>% rename("Color" = "color") #all other data sets use the capitalized versi
```

Creating columns indicating the categories of each item:

```
highfashion <- highfashion %>% mutate(Category = "High Fashion")
depop <- depop %>% mutate(Category = "Depop")
nordstrom <- nordstrom %>% mutate(Category = "Nordstrom")
zara <- zara %>% mutate(Category = "Zara")
hm <- hm %>% mutate(Category = "H&M")
shein <- depop %>% mutate(Category = "Shein")
```

Creating Contingency Tables

Creating a contingency table for the High Fashion and Depop datasets based on the frequncy of colors by category:

```
high_depop <- full_join(depop, highfashion, by = c("Category", "Color")) #joining the depop and high fa
high_depop <- high_depop %>% distinct() #ensuring there are no duplicate entries

table.high_depop <- high_depop %>% select(c("Category", "Color")) %>% table() #creating a contingency t
table.high_depop
```

```
##               Color
## Category       beige black blue brown denim gray green grey multicolored navy
##   Depop            2    13    2     1     5    1     3    1           5    1
##   High Fashion     4     6    6     4     0    0     3    0          14    0
##               Color
## Category       off-white orange pastels pink purple red silver white white
##   Depop                0      0       1    2      1   2      0     7      1
##   High Fashion         3      2       0    3      0   0      1     9      0
##               Color
## Category       yellow
##   Depop             1
##   High Fashion      4
```

High Fashion and Depop: Tests & Visualizations

Sampling is done at random & outcomes are mutually exclusive

Null hypothesis (H0): There is no association between the two variables.
Alternative hypothesis (HA): There is an association between the two variables.

Fisher's Exact Test:
p-value of 0.01182
0.01182 < alpha = 0.05
Null hypothesis can be rejected. Therefore, there is an association between the colors found in High Fashion and Depop products.

Cramer's V:
Cramer's V measures the association between two variables with a result of 0 indicating no association and a result of 1 indication a perfect association. Cramer's V for the association between colors found in High Fashion and Depop products is 0.5332. This indicates a somewhat strong association.

Conducting the tests:

```
degrees_of_freedom <- min(nrow(table.high_depop)-1, ncol(table.high_depop)-1)
degrees_of_freedom
```

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

```
fisher.test(high_depop$Category,high_depop$Color)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  high_depop$Category and high_depop$Color
## p-value = 0.01096
## alternative hypothesis: two.sided
```

```
cramerV(high_depop$Category,high_depop$Color)
```

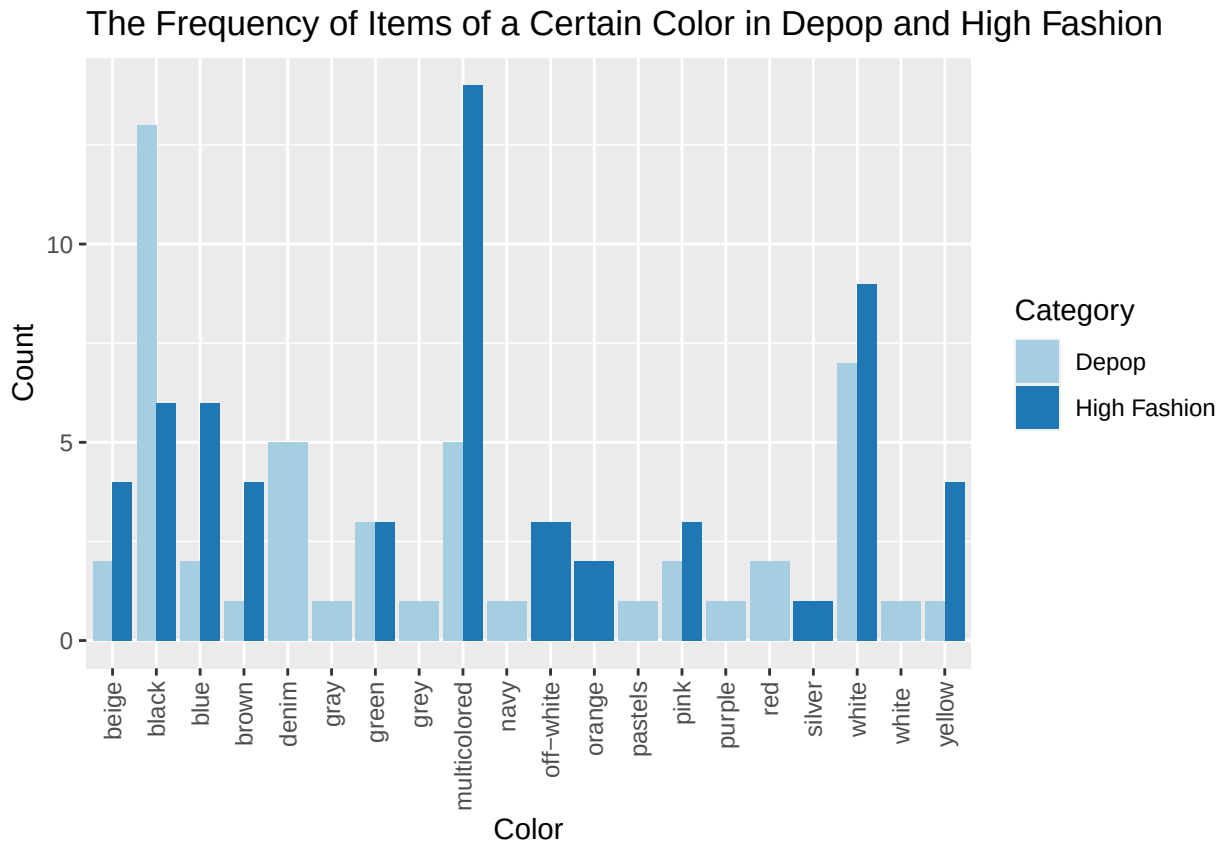
```
## Cramer V
##   0.5435
```

Note: This visualization (and analysis overall) may be improved by grouping together certain colors into more general categories. ie: "off-white" and "white" can be combined into just "white" or "denim" and "blue" can be combined into just "blue".

Visualizing the distributions by plotting:

```
color.high_depop <- high_depop %>% select(c("Category", "Color"))

ggplot(color.high_depop, aes(x=Color, fill= Category)) +
  geom_bar(position = "dodge")+
  scale_fill_brewer(palette = 'Paired',labels = c('Depop', 'High Fashion')) +
  scale_x_discrete(guide = guide_axis(angle = 90))+
  ylab("Count")+
  ggtitle("The Frequency of Items of a Certain Color in Depop and High Fashion")
```



Fifth Experiment:

Are trousers becoming more popular than other styles of bottoms?

Hypothesis: Trousers are seen more in current fashion trends than other types of bottoms

Process: Conduct a two-sided z-test to compare the proportion of trousers compared to other styles of bottoms.

Data Manipulation:

```
# filter out only the rows with pants or shorts

pinterest_pants <- pinterest %>%
  filter(Article %in% c("pants", "shorts"))

hm_pants <- hm %>% # none
  filter(Article %in% c("pants", "shorts"))

shein_pants <- shein %>% # none
  filter(Article %in% c("pants", "shorts"))

depop_pants <- depop %>%
  filter(Article %in% c("pants", "shorts"))

highfashion_spring_pants <- highfashion_spring %>%
```

```
  filter(article %in% c("pants", "shorts"))

nordstrom_pants <- nordstrom %>%
  filter(Article %in% c("pants", "shorts"))

zara_pants <-  zara %>% # none
  filter(Article %in% c("pants", "shorts"))

head(pinterest_pants) # 12 trousers
```

```
##   Article   Fit Material Rise   Embellishment Length Sleeve Neckline Pockets
## 1   pants loose          high        trousers                              no
## 2   pants  lose           mid           cargo                             yes
## 3   pants loose           low parachute/cargo                             yes
## 4   pants loose           low           cargo                             yes
## 5   pants loose          high        trousers                              no
## 6   pants loose          high        trousers                              no
##       Color                   Search
## 1 dark grey trendy outfits for women
## 2     cream trendy outfits for women
## 3     cream           trendy outfits
## 4 dark grey           trendy outfits
## 5       tan           trendy outfits
## 6       tan                  outfits
```

```
highfashion_spring_pants <- rename(highfashion_spring_pants, Article = article)
head(highfashion_spring_pants) # 1 trouser
```

```
##   Article      fit material embellishment length sleeve neckine pockets   brand
## 1   pants    loose     wool       pleated   long   <NA>    <NA>       y The Row
## 2   pants straight    denim      slim-fit   long   <NA>    <NA>       y   Alaia
##   color      rise cropped
## 1 brown high-rise    <NA>
## 2  blue high-rise       n
```

```
head(nordstrom_pants) # 1 trouser
```

```
##   Article     Fit     Material Rise..pants.skirts. Embellishment
## 1  shorts relaxed        denim                mid        cutoff
## 2   pants  fitted         <NA>               high          <NA>
## 3  shorts relaxed linen/cotton               high          <NA>
##   Length..dress.skirts. Sleeve.type Neckline Pockets                     Color
## 1                  <NA>        <NA>     <NA>     yes                      blue
## 2                  <NA>        <NA>     <NA>     yes gray/green/white/red/blue
## 3                  <NA>        <NA>     <NA>     yes                pink/gray
##
## 1                                      https://www.nordstrom.com/s/kut-from-the-kloth-gidget-denim
## 2 https://www.nordstrom.com/s/wit-and-wisdom-absolution-high-waist-ankle-skinny-pants-regular-petite
## 3                                                https://www.nordstrom.com/s/madewell-clean-linen-c
##    Category
## 1 Nordstrom
## 2 Nordstrom
## 3 Nordstrom
```

Data Organization:

```
# make data frame with non-trousers

pinterest_other <- pinterest_pants %>%
  filter(is.na(Embellishment) | Embellishment != "trousers")

highfashions_other <- highfashion_spring_pants %>%
  filter(material ==  "denim")

nordstrom_other <- nordstrom_pants %>%
  filter(Material != "linen/cotton")


other_pants <-  bind_rows(pinterest_other, highfashions_other)
other_pants <- bind_rows(other_pants, nordstrom_other)
other_pants$Trouser <- "no" # make new row for boolean variable trousers and assign all to "no"
```

```
# make data frame with all trousers
pinterest_trousers <- pinterest_pants %>% filter(Embellishment ==  "trousers")
highfashions_trousers <- highfashion_spring_pants %>% filter(embellishment ==  "pleated")
nordstrom_trousers <- nordstrom_pants %>% filter(Material == "linen/cotton")


trousers <-  bind_rows(pinterest_trousers, highfashions_trousers)
trousers <- bind_rows(trousers, nordstrom_trousers)
trousers$Trouser <- "yes" # make new row for boolean variable trousers and assign all to "yes"
```

```
# combine both trouser and non-trouser data frames
all_pants <-  bind_rows(trousers, other_pants)
```

Write hypotheses & check conditions:

Null Hypothesis: prop1 = prop2 Alternative Hypothesis: prop1 not equal to prop2

We can assume that the data is approximately normal since n $>$ 30. The data was randomly collected.

Set alpha at 0.05

```
# calculate sample size and proportion
sample_size = 40
sample_prop = (14/40)
hypoth_prop = 0.5 # where proportion of trousers are equal to other types of pants
```

```
# perform z test
test <- prop.test( x = sample_prop, n = sample_size, p = hypoth_prop )
test
```

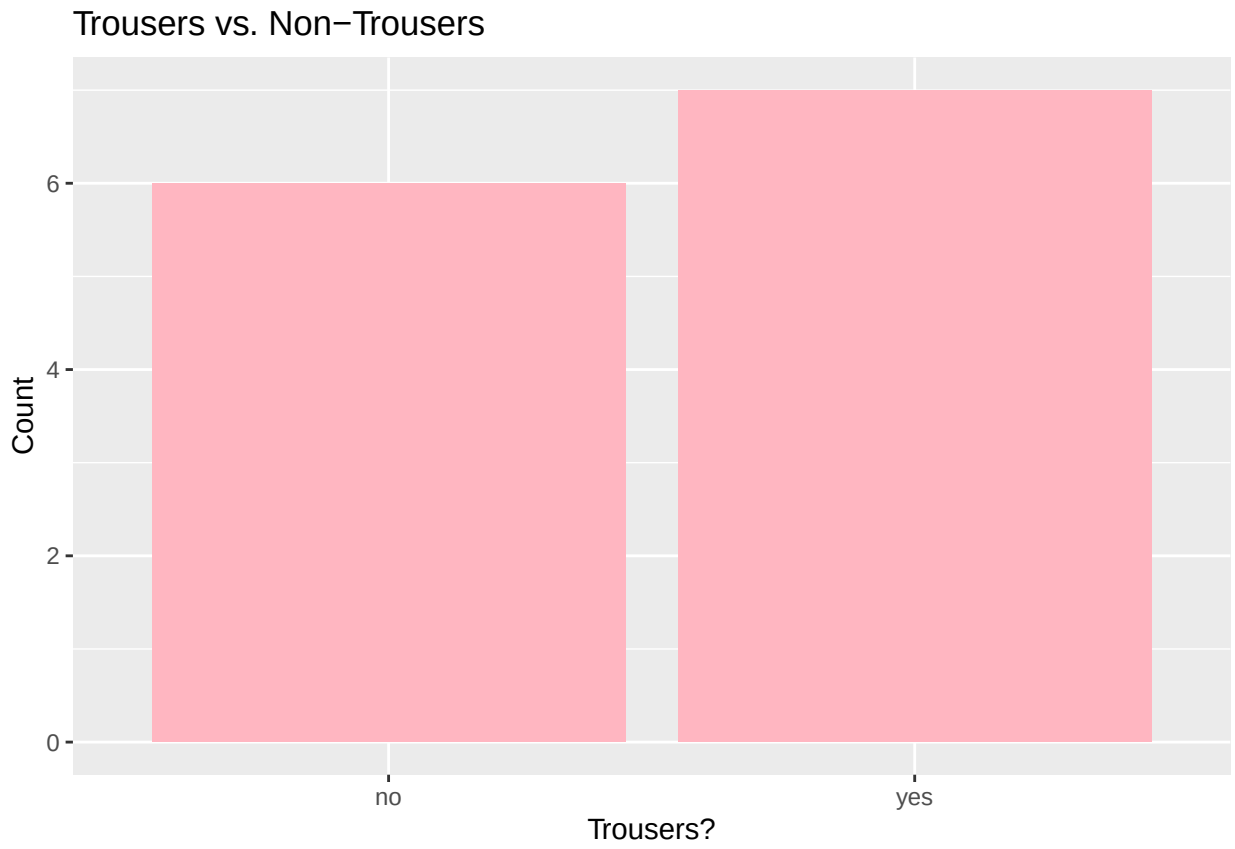
```
##
##  1-sample proportions test with continuity correction
##
## data:  sample_prop out of sample_size, null probability hypoth_prop
## X-squared = 36.672, df = 1, p-value = 1.398e-09
```

```
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.0000000 0.1230495
## sample estimates:
##       p
## 0.00875
```

Since the p-value is less than alpha = 0.05, there is enough evidence to reject the null hypothesis. This means that the proportion of trousers is not equal to the proportion of other pants. Based on this test, we can conclude that trousers might be overtaking the other styles of pants in popularity among various retailers.

Visualization:

```
ggplot(all_pants, aes(x = Trouser)) +
  geom_bar(fill = "lightpink") +
  labs(title = "Trousers vs. Non-Trousers",
       x = "Trousers?",
       y = "Count")
```



Trousers Conclusion:

We were specifically curious about investigating this because it is a trend that everyone on our team has witnessed in our day-to-day life. That is one of the special things about the way we conducted our research: we got to investigate whatever we wanted to. We had a suspicion that trousers are the most popular type of pants for women at the moment, and our data reaffirmed our hypothesis

From Our Team:

We have thoroughly enjoyed exploring our interests in fashion and data science throughout this research project, and we have collectively learned so much. Although still at an elementary level, this project was an opportunity to follow-through with every step of data science research, from data collection to forming conclusions. We would be interested in exploring these topics further in the future, and as our skills develop, we will be able to discover even more interesting results.

Thank you for your interest in our work,

The Quetzal Team:

Esha Chakrabarty, Hannah Wen, Jamie Chac, Libby Amir, & Pearl Vishen