Shopify DS

Victoria Okoro

1/17/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.4
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

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

```
library(ggplot2)
library(plyr)
```

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(dplyr)
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, mutate, rename, summarise
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(cowplot)
```

```
df <- read.csv("~/Documents/Shopify/Copy of 2019 Winter Data Science Intern Challenge Data Set - Sheet1
```

```
summary(df$order_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      90     163     284    3145     390  704000
```

```
median(df$order_amount)
```

```
## [1] 284
```

```
df$total_items <- sort(df$total_items)

bp <- ggplot(data = df, aes(x = total_items)) +
  geom_bar(fill = "purple") +  ggtitle("Distribution of Items Bought.") + xlim(0,8) +
  xlab("Total Items")

ggplotly(bp)
```

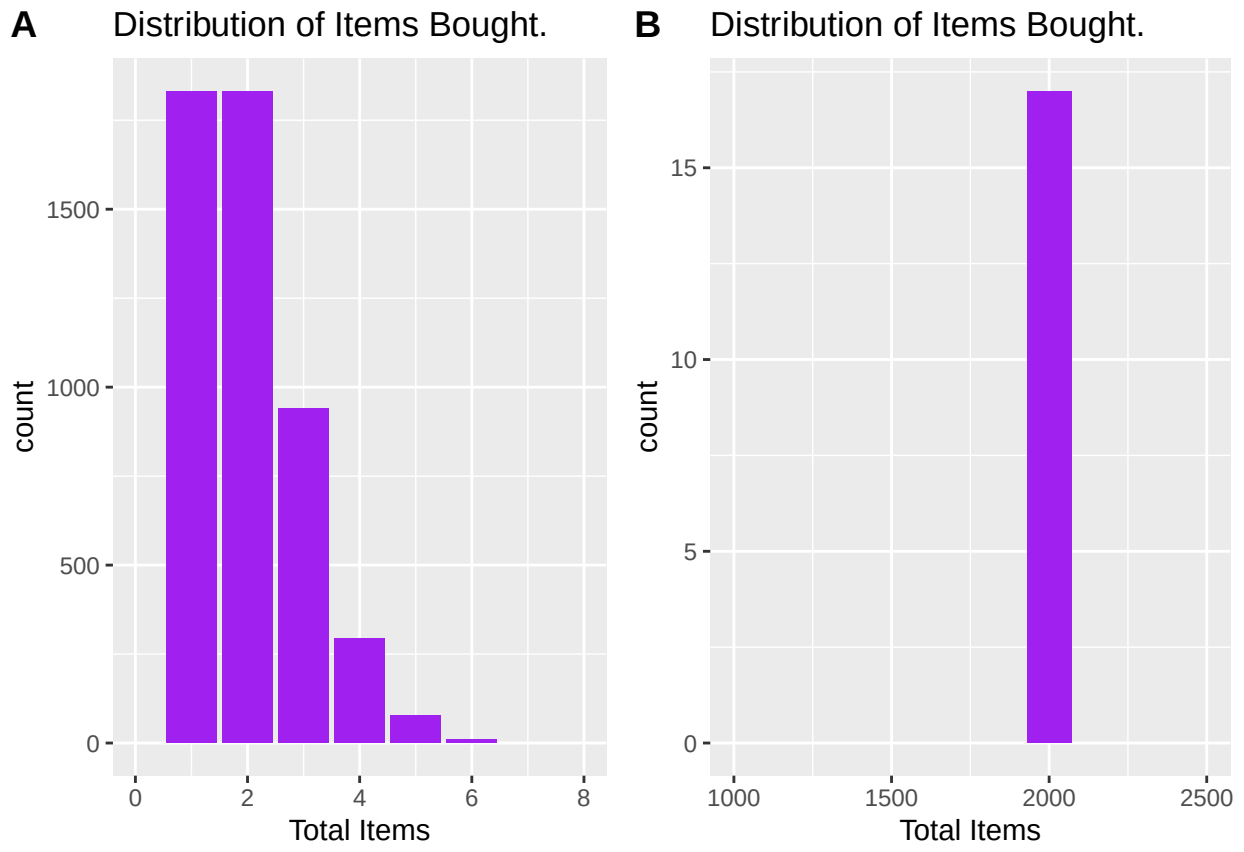
```
## Warning: Removed 17 rows containing non-finite values (stat_count).
```

```
df2 <- df[4984:5000, ]
df2$total_items <- sort(df2$total_items)

bp2 <- ggplot(data = df2, aes(x = total_items)) +
  geom_bar(fill = "purple", width = 140) +
  ggtitle("Distribution of Items Bought.") + xlim(1000,2500) +
  xlab("Total Items")

ggplotly(bp2)
```

```
bp_grid <- plot_grid(bp, bp2, labels = "AUTO")
```

```
## Warning: Removed 17 rows containing non-finite values (stat_count).

## Warning: Removed 1 rows containing missing values (geom_bar).
```

```
bp_grid
```



##Summary:

The median is the best indicator for how much a customer usually spends on sneakers given they are a relatively affordable, rather than using the mean. For my visualization I illustrated the frequency of the amount of items customers bought. Majority of customers bought between 1 to 3 items. Only a small amount of customers bought 2000 items. The count is:

1 -> 1830 2 -> 1832 3 -> 941 4 -> 293 5 -> 77 6 -> 9 2000 -> 17