Shopify_Challenge

Ankita Sarkar

10/30/2021

Reading the data set:

```r
file <- read.csv('Sheet1.csv')
```

Creating a data frame to work with it

```r
df <- data.frame(file)
```

Getting the summary of the data

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

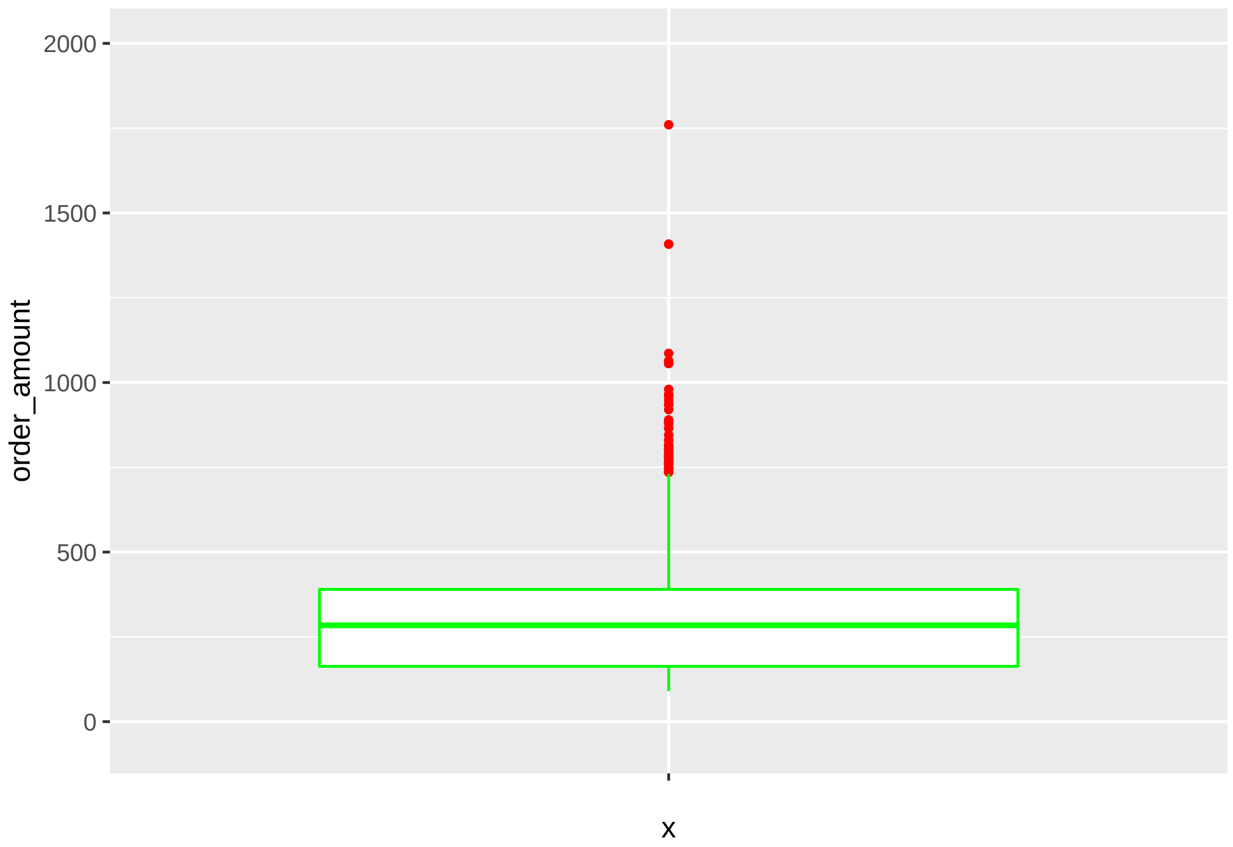
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##      90     163     284    3145     390  704000
```

Clearly the maximum order_amount for the sneakers is way too long. This is the reason we have got higher average value i.e. $3145. While the minimum order_amount is 90.

Plotting a boxplot for the data

```r
library(ggplot2)

ggplot(data = df) +
  aes(x = '', y = order_amount) +
  geom_boxplot(outlier.colour = 'red', outlier.size = 1, colour = 'green') +
  coord_cartesian(ylim = c(-50, 2000))
```



We have plotted the order_amount with a box plot (in green). The y axis is set between -50 and 2000. The outliers present in the dataset are shown in red. The median is at 284, first quantile is at 163 and the third quantile is at 390.

Printing out outlier values

```
out <- boxplot.stats(df$order_amount)$out
out
```

```
##   [1] 704000 704000    780    765  25725    780    765    780    780  51450
##  [11]  51450  51450 704000    830  51450    748 154350    772    804    815
##  [21]    885   1056    784  25725 704000    815    885  25725  25725    935
##  [31]  77175 704000   1760   1408  25725  25725 704000  25725   1408    765
##  [41]    736  51450 704000    960 704000    800    804    800    865    745
##  [51]    830    880    920    765    774    790    784 704000  25725 704000
##  [61]    948    845    760    745  51450 102900    965  51450  51450  25725
##  [71]    935  77175    780  77175    805  25725  51450  51450 704000  77175
##  [81]  25725    830 704000   1056    890    980  25725  51450    760  25725
##  [91]  51450    748    786 704000  77175    736    805  25725   1056    736
## [101]    935   1086    736  51450  77175  25725    816    810    740  25725
## [111] 704000  51450   1064  77175    780  51450  51450  77175    735  25725
## [121]    760    880    780    748    748  25725    748    800 704000    780
## [131]  77175    960 704000    790 704000    760  25725    765    880    865
## [141]    772
```

Printing out the indexes containing outliers

```
out_ind <- which(df$order_amount %in% c(out))
out_ind
```

```
##   [1]   16   61  100  137  161  220  223  260  265  491  494  512  521  523  618
##  [16]  652  692  738  743  772  880  939  995 1057 1105 1124 1151 1194 1205 1257
##  [31] 1260 1363 1365 1368 1385 1420 1437 1453 1472 1485 1504 1530 1563 1564 1603
##  [46] 1629 1765 1924 1947 1949 1963 2033 2040 2044 2128 2137 2141 2154 2271 2298
##  [61] 2308 2354 2387 2390 2453 2493 2495 2496 2513 2549 2561 2565 2671 2691 2758
##  [76] 2774 2819 2822 2836 2907 2923 2968 2970 2988 3074 3078 3086 3102 3118 3152
##  [91] 3168 3203 3253 3333 3404 3429 3439 3441 3514 3518 3533 3539 3610 3706 3725
## [106] 3781 3866 3928 3967 4041 4057 4080 4142 4193 4296 4312 4413 4421 4491 4506
## [121] 4513 4524 4555 4575 4581 4585 4597 4620 4647 4712 4716 4848 4869 4871 4883
## [136] 4906 4919 4928 4953 4959 4981
```

There are no specific way to handle outliers in data. We should not really ignore outliers which is quite common in real life dataset. One approach we can take to handle the outlier problem in the present dataset is to replace them with the value of median. As median does not get affected with the presence of outliers unlike mean.

We see the median for order_amount is 284. So we would change all the rows holding excessive large amount with this median and process again.

```
df[out_ind, ] <- median(df$order_amount)
```

If we check the summary of the data now, it seems to solve our problem. The minimum as is at 90 and maximum at 730 while we get the mean at 293.4.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    90.0   163.0   284.0   293.4   374.0   730.0
```