Shopify Challenge

Richard Ye

14/09/2021

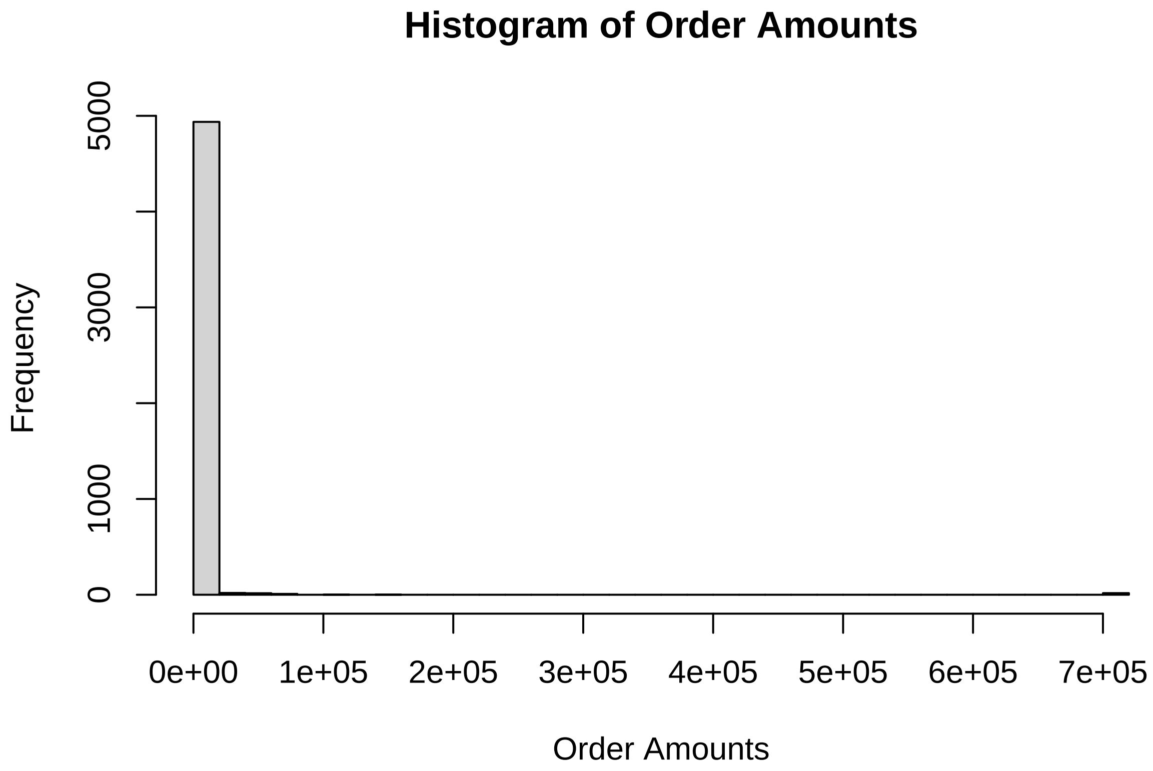
```
library(tidyverse)
library(ggplot2)
library(readxl)
data = read_xlsx(file.path(getwd(), "2019 Winter Data Science Intern Challenge Data Set.xlsx"))
```

Question 1

On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of $3145.13. Given that we know these shops are selling sneakers, a relatively affordable item, something seems wrong with our analysis.

(a) **Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.**
Because we're just looking at the average order value, (ie the mean of the order amount), we are leaving ourselves open for outlying data points to greatly skew our results. Let's take a look at how order amounts are distributed:

```
hist(data$order_amount, breaks = 50, main = "Histogram of Order Amounts", xlab = "Order Amounts")
```



From this distribution, we can see that though the majority of the order amounts lie under $100,000, it seems we have an order over $700,000.

```
max(data$order_amount)
```

```
## [1] 704000
```

```
sort(data$order_amount, decreasing = TRUE) %>% head(20)
```

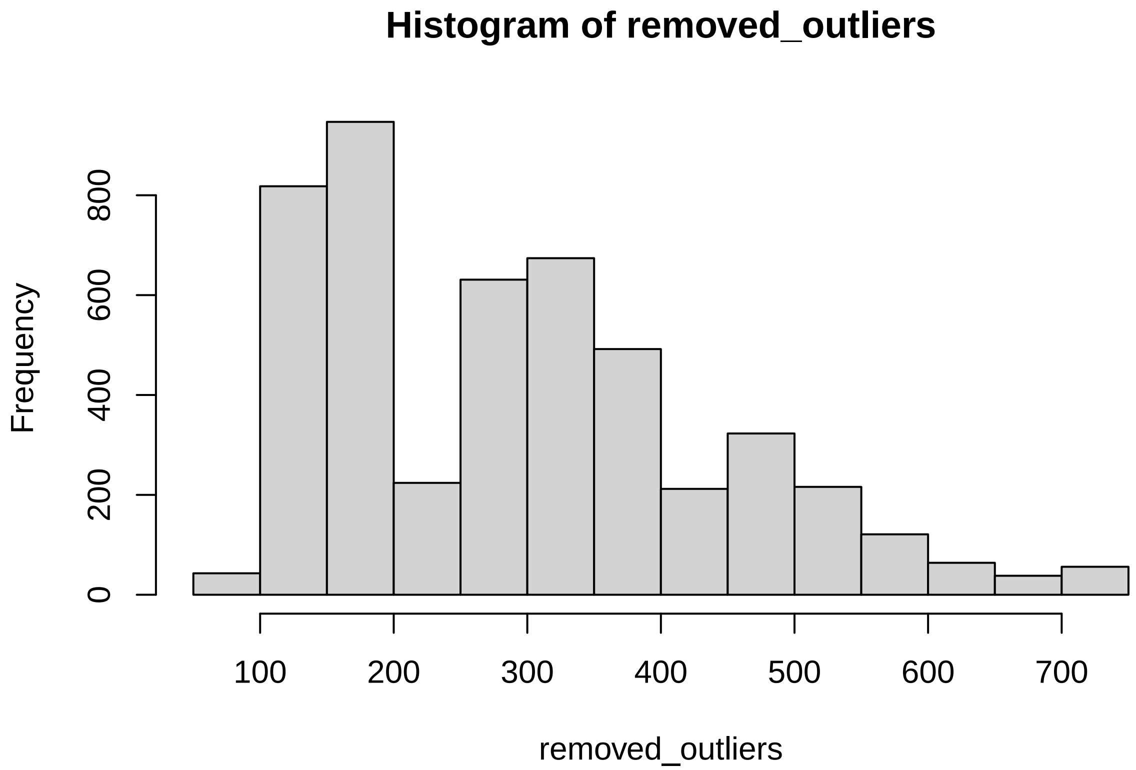
```
##  [1] 704000 704000 704000 704000 704000 704000 704000 704000 704000 704000
## [11] 704000 704000 704000 704000 704000 704000 704000 154350 102900  77175
```

```
sort(data$order_amount, decreasing = TRUE) %>% tail(20)
```

```
##  [1] 94 94 90 90 90 90 90 90 90 90 90 90 90 90 90 90 90 90 90 90
```

The highest order amount is indeed 704,000. We can see that there are other large order amounts around 100,000 and 70,000, which we can tell from the histogram above make up very few of the actual orders. Lets take a look at the distribution without all outliers:

```
removed_outliers =data$order_amount[!data$order_amount %in% boxplot.stats(data$order_amount)$out]
hist(removed_outliers)
```



We can see that without outliers, order amounts just under $200 happen the most often.

A better way to evaluate this data, if we wanted to examine the average order amount would be to consider the median and modal value. The median is good because this is a skewed data set with outliers, and the median is less affected. The mode however gives us the most frequent value which in this case would lie somewhere between $150 and $200.

```
median(data$order_amount)
```

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

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(data$order_amount)
```

```
## [1] 153
```

We obtain a median order amount of $284, and mode order amount of $153 both of which makes much more sense.

(b) What metric would you report for this dataset? The median would best represent the average order value, since it is quite uncommon for orders over $100,000 to occur and it tells us what order amount most accurately represents the average. I chose this over the mode because it is less effected by the skewness of our data and outliers, and still gives us a representative value of the central location of the data. The mode makes intuitive sense and completely ignores the outliers, and skewness of the data, but the drawbacks to using this method include when the more frequent value is far from all other values and when two order amounts occur equally as frequently.

(c) What is its value? The value of the median order amount is $284.

Question 2: For this question you'll need to use SQL. Follow this link to access the data set required for the challenge. Please use queries to answer the following questions. Paste your queries along with your final numerical answers below.

How many orders were shipped by Speedy Express in total?

```sql
SELECT COUNT(*) FROM [Orders] as o
LEFT JOIN [Shippers] as sh ON o.ShipperID = sh.ShipperID
Where sh.ShipperName = "Speedy Express"
```

Ans: 54

What is the last name of the employee with the most orders?

```sql
SELECT e.LastName, COUNT(*) as Orders FROM [Orders] as o
LEFT JOIN [Employees] as e ON o.EmployeeID = e.EmployeeID
GROUP BY e.LastName, e.FirstName
ORDER BY Orders desc
LIMIT 1
```

Ans: Peacock with 40 orders

What product was ordered the most by customers in Germany?

```sql
SELECT p.ProductName, SUM(od.Quantity) as Quantity, c.Country FROM [Orders] AS o
LEFT JOIN [OrderDetails] as od ON o.OrderID = od.OrderID
LEFT JOIN [Customers] as c on o.CustomerID = c.CustomerID
LEFT JOIN [Products] as p on od.ProductID = p.ProductID
WHERE c.Country = "Germany"
GROUP BY od.ProductID
ORDER BY Quantity DESC
LIMIT 1
```

Ans: Boston Crab Meat at 160 items ordered over 4 orders. This definition of most is based on the assumption we want the item with the highest total quantity ordered by customers in Germany.