Shopify Intern Challenge

Nick Finn

September 20, 2021

Question 1

```
# Load the data set utilizing the readxl library for xlsx files.
library("readxl")
shoe.data <- read_excel("C:/Users/sedky/Documents/Job Stuff/Shopify/2019 Winter Data Science Intern Cha
```

a)

```
# Diagnostics

# AOV
mean(shoe.data$order_amount)
```

```
## [1] 3145.128
```

```
# Range values Order amounts
summary(shoe.data$order_amount)
```

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

```
# Total items sold
summary(shoe.data$total_items)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    1.000    1.000    2.000    8.787    3.000 2000.000
```

The reason that this value would seem high is due to the fact that Average order value is calculated as average money spent per order, and in a lot of orders a ton of shoes can be bought for really high amounts. This can be seen in the summaries, where the difference between min and max values for Order Amount and Total Items is huge, thus effectively skewing the average order value to a really high amount. This of course leads to a massive skew in the data, thus raising the average value way above the median, which in this case would be a better metric of central value.

b)

As said previously, a better metric of evaluation in this case would be to simply just use the Median Order Value. The reason for this is because we have a highly skewed distribution in terms of Order Amounts, this of course due to what I mentioned previously with the fact that there are extremely high Order Amounts way above the median. In situations with skewed data like this, it's a lot better to utilize the median as apposed to the average to avoid the effect of extreme outlying values, like the few $704,000 Order Amounts.

c)

```
median(shoe.data$order_amount)
```

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

The Median order Value is $284, which is a way more reasonable metric to utilize in an analysis where the majority of customers will only be buying around 1-3 shoes based on the summaries.

Question 2

a)

SELECT COUNT(*) FROM Orders;

Result: 196

b)

SELECT LastName FROM Employees JOIN Orders ON Employees.EmployeeID = Orders.EmployeeID GROUP BY LastName ORDER BY COUNT(*) DESC LIMIT 1;

Result: Peacock

c)

SELECT ProductName FROM Products JOIN OrderDetails ON Products.ProductID = OrderDetails.ProductID JOIN Orders ON OrderDetails.OrderID = Orders.OrderID JOIN Customers ON Orders.CustomerID = Customers.CustomerID WHERE Customers.Country == "Germany" GROUP BY ProductName ORDER BY COUNT(*) DESC LIMIT 1;

Result: Gorgonzola Telino