Shopify_Data_Challenge

Yujun Mu

14/09/2021

Q1 R

```
options(scipen=999)
library(mice)
library(ggplot2)
library(corrplot)
library(EnvStats)
library(dplyr)

# Mode function to get the mode stats
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

# read data
data <- readxl::read_xlsx("data.xlsx")
summary(data)
```

```
##     order_id       shop_id          user_id       order_amount
##  Min.   :   1   Min.   :  1.00   Min.   :607.0   Min.   :    90
##  1st Qu.:1251   1st Qu.: 24.00   1st Qu.:775.0   1st Qu.:   163
##  Median :2500   Median : 50.00   Median :849.0   Median :   284
##  Mean   :2500   Mean   : 50.08   Mean   :849.1   Mean   :  3145
##  3rd Qu.:3750   3rd Qu.: 75.00   3rd Qu.:925.0   3rd Qu.:   390
##  Max.   :5000   Max.   :100.00   Max.   :999.0   Max.   :704000
##   total_items       payment_method       created_at
##  Min.   :   1.000   Length:5000        Min.   :2017-03-01 00:08:09
##  1st Qu.:   1.000   Class :character   1st Qu.:2017-03-08 07:08:04
##  Median :   2.000   Mode  :character   Median :2017-03-16 00:21:20
##  Mean   :   8.787                      Mean   :2017-03-15 22:20:37
##  3rd Qu.:   3.000                      3rd Qu.:2017-03-23 10:39:57
##  Max.   :2000.000                      Max.   :2017-03-30 23:55:35
```

```
# AOV = Total Revenue / Total Number of orders
sum(data$order_amount) / sum(data$total_items)
```

```
## [1] 357.9215
```

```
# AOV by different stores
aov_byshop <- data %>%
  group_by(shop_id) %>%
```

```
  summarise(sum(order_amount)/sum(total_items))

# rename the col names
colnames(aov_byshop) <- c("shop_id","AOV")

# Calculate the mean, median, and mode for order values based on different sh
ops
mean(aov_byshop$AOV)

## [1] 407.99

median(aov_byshop$AOV)

## [1] 153

Mode(aov_byshop$AOV)

## [1] 153
```

a. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.

In the current calculation, the AOV is the simple average amount with all the orders regardless of the items purchased per order. Therefore, in the next step, I will try to add on that information.

b. What metric would you report for this dataset? c. What is its value?

I recalculate the AOV by taking the total items into account. The value decreases to around $358 but still consider high for shoes. Then, I calculate the AOV for each store. The interesting find is that one particular store, #78, has an AOV of around $25,725. Therefore, it influences the average. As the final step, I output the mean, median, and mode for the average order value. It turns out the median and mode are both about $153, which is in practice, and the mean is about $407.99 high. This is surely influenced by store 78.

Q2 SQL

a. How many orders were shipped by Speedy Express in total?

Answer: *54*

```
SELECT
COUNT(ShipperID)
FROM Orders
WHERE ShipperID == 1
```

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

Answer: *Peacock with 40*

```sql
CREATE VIEW EO AS
SELECT Orders.OrderID, Orders.EmployeeID, Employees.LastName
FROM Orders
JOIN Employees
ON Orders.EmployeeID = Employees.EmployeeID

SELECT LastName, COUNT(*)
FROM EO
GROUP BY LastName
ORDER BY COUNT(*) desc
```

c. What product was ordered the most by customers in Germany?

Answer: 160 Boston Crab Meat

```sql
CREATE VIEW C AS
SELECT Customers.CustomerID, Customers.Country, Orders.OrderID
FROM Orders
JOIN Customers
ON Orders.CustomerID = Customers.CustomerID

CREATE VIEW D AS
SELECT OrderDetails.ProductID, OrderDetails.Quantity, C.OrderID, C.Country
FROM C
JOIN OrderDetails
ON C.OrderID = OrderDetails.OrderID

CREATE VIEW V AS
SELECT D.ProductID,SUM(Quantity) as Total_Q
FROM D
WHERE COUNTRY == 'Germany'
GROUP BY ProductID

CREATE VIEW P AS
SELECT Products.ProductID, Products.ProductName,V.Total_Q
FROM V
JOIN Products
ON Products.ProductID = V.ProductID

SELECT MAX(Total_Q), ProductName,ProductID
FROM P
```