2022 Winter Data science challenge

Khusanbek Mukhammad Azim

12/09/2021

Question 1

We first read in the data from the csv and check that we have all the columns.

```
sneaker <- read_csv("2019_challenge_set.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   order_id = col_double(),
##   shop_id = col_double(),
##   user_id = col_double(),
##   order_amount = col_double(),
##   total_items = col_double(),
##   payment_method = col_character(),
##   created_at = col_character()
## )
```

(a)

Lets take a look at our data.

```
sneaker
```

```
## # A tibble: 5,000 x 7
##    order_id shop_id user_id order_amount total_items payment_method created_at
##       <dbl>   <dbl>   <dbl>        <dbl>       <dbl> <chr>          <chr>
##  1        1      53     746          224           2 cash           2017-03-13 ~
##  2        2      92     925           90           1 cash           2017-03-03 ~
##  3        3      44     861          144           1 cash           2017-03-14 ~
##  4        4      18     935          156           1 credit_card    2017-03-26 ~
##  5        5      18     883          156           1 credit_card    2017-03-01 ~
##  6        6      58     882          138           1 credit_card    2017-03-14 ~
##  7        7      87     915          149           1 cash           2017-03-01 ~
##  8        8      22     761          292           2 cash           2017-03-08 ~
##  9        9      64     914          266           2 debit          2017-03-17 ~
## 10       10      52     788          146           1 credit_card    2017-03-30 ~
## # ... with 4,990 more rows
```

Notice that for order id 4 and 5 come from the same shop based on the shop id. I suspect that the average of the order_amount column was taken without accounting for the fact that the same shop can have multiple order ids.

Let us check this assumption by calculating the average of the orders.

```
mean(sneaker$order_amount)
```

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

Here we get that the average order value is $3145.13 as first assumed.

(b)

We can group all shop purchases together and then take the median of all their purchases. Since a couple of the shops have large orders, which skews the average, taking the median or excluding the outliers and then taking the average would also work.

Below we see what the AOV for each shop is.

```
sneaker %>%
  group_by(shop_id) %>%
  summarize(shop_AOV = mean(order_amount)) -> shop_order

shop_order
```

```
## # A tibble: 100 x 2
##    shop_id shop_AOV
##  *   <dbl>    <dbl>
##  1       1     309.
##  2       2     174.
##  3       3     305.
##  4       4     259.
##  5       5     290.
##  6       6     384.
##  7       7     218
##  8       8     241.
##  9       9     234
## 10      10     332.
## # ... with 90 more rows
```

(c)

Next we find the median of the shop AOV to get to our answer.

```
median(shop_order$shop_AOV)
```

```
## [1] 308.8898
```

The median order value is $308.89.

```
mean(shop_order$shop_AOV[1:41]) + mean(shop_order$shop_AOV[43:77]) + mean(shop_order$shop_AOV[79:100])
```

```
## [1] 902.4913
```

Or if we instead exclude shop ids 42 and 78 and calculate the average we get a value of $902.49.

Question 2

(a)

SELECT ShipperID, count(*) from Orders

GROUP BY ShipperID

Number of Records: 3

ShipperID	count(*)
1	54
2	74
3	68

Speedy Express has Shipper ID of 1 shipped 54 orders.

(b)

SELECT OrderDetails.OrderID, OrderDetails.Quantity, Orders.OrderID, Orders.EmployeeID, Employees.EmployeeID, Employees.LastName

FROM OrderDetails **JOIN** Orders

ON OrderDetails.OrderID = Orders.OrderID

JOIN Employees

ON Employees.EmployeeID = Orders.EmployeeID

GROUP BY LastName

ORDER BY Quantity DESC

Number of Records: 9

OrderID	Quantity	EmployeeID	LastName
10258	50	1	Davolio
10289	30	7	King
10265	30	2	Fuller
10255	20	9	Dodsworth
10262	12	8	Callahan
10248	12	5	Buchanan
10250	10	4	Peacock
10249	9	6	Suyama
10251	6	3	Leverling

Davolio had the most orders.

(c)

SELECT Customers.Country, Products.ProductName

FROM Customers

INNER JOIN Orders **ON** Customers.CustomerID = Orders.CustomerID

INNER JOIN OrderDetails **ON** Orders.OrderID = OrderDetails.OrderID

INNER JOIN Products **ON** OrderDetails.ProductID = Products.ProductID

GROUP BY Country

Number of Records: 21

Country	ProductName
Argentina	Tofu
Austria	Chang
Belgium	Sir Rodney's Marmalade
Brazil	Tofu
Canada	Carnarvon Tigers
Denmark	Geitost
Finland	Queso Cabrales
France	Gustaf's Knäckebröd
Germany	Boston Crab Meat
Ireland	Chang
Italy	Guaraná Fantástica
Mexico	Sir Rodney's Scones
Norway	Guaraná Fantástica
Poland	Gorgonzola Telino
Portugal	Raclette Courdavault
Spain	Teatime Chocolate Biscuits
Sweden	Chang
Switzerland	Guaraná Fantástica
UK	Aniseed Syrup
USA	Jack's New England Clam Chowder
Venezuela	Schoggi Schokolade

The most ordered product by customers from Germany was Boston Crab Meat.