# Fall 2021 Data Science Intern Challenge

Please complete the following questions, and provide your thought process/work. You can attach your work in a text file, link, etc. on the application page. Please ensure answers are easily visible for reviewers!

**Question 1:** Given some sample data, write a program to answer the following: [click here to access the required data set](https://docs.google.com/spreadsheets/d/16i38oonuX1y1g7C_UAmiK9GkY7cS-64DfiDMNiR41LM/edit#gid=0)

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.

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

The wrong result was calculated by adding all the order amount and divided by the number of records(15725640/ 5000 = 3125.13). By definition, AOV is calculated using revenue divided by number of orders and it is used to tracks the average dollar amount spent every time a customer places an order. When look at the data, I noticed there are some outliers exists that cause the value way much higher than the nomal sneaker prices. To be specific, the userid 607 had purchased 2000 pairs of shoes for each order and had made 17 times order during the 30-day window. Other orders normally had amount from 1-8. Since the records contain 100 shops, I first calculated AOV for every shop by writing sql queries and results table are showed in appendix(Table A). Once we separate by different shops the AOV values become more reasonable than the original except for shop 42 and 78, which contains outliers with regard to the amount of items in one order. If we take out of shop 42 and 78 and recalculate the AOV for the other shops, the AOV value is 300.16. On the other hand, I also wrote a sql query to check the price of the shoe in different shops and I noticed the shoe price in shop 78 is very irregular($25725) and I have think of two possible reasons: 1. The record missed a decimal point, the price is actually $257.25. If that is the case, I would not regard records from shop 78 as outliers and recalculate the AOV to be 302; 2. There is no problem with the shoe price since it is luxury brand. If that is the case, nothing is needed to be changed.

1. What metric would you report for this dataset?

Besides AOV, I also used average basket value, purchase frequency and customer value to get a more complete picture of the profitability of sneaker business.

**Average basket value**—The metrics measures the average amount of items sold per order. It is calculated using the total number of units sold divided by the number of orders. When combined with AOV, average basket value can offer a more complete picture of sales trends.

**Purchase frequency**--Purchase frequency is a metric that shows the average number of orders placed by each customers. It is calculated by choosing a particular timeframe(in this question, 30-day) and then dividing the number of orders by the number of unique customers.

**Customer value**-- It is a measure of how much revenue a customer brings to the business during a specific timeframe. It is calculated by multiplying AOV by purchase frequency.

1. What is its value?

In terms of AOV, I removed the two stores(42 and 78) that contain ourlier values for price and order amount and recalculated it and get the value of 300.16. The AOV for store 42 and 78 are I also generated AOV for every store and the results are shown in appendix.

In terms of average basket value, Except for shop 42 with basket size of 668, all the other shop had basket size of 2.

In terms of purchase frequency, all the shops had the frequency of 1. If including all the shops, the purchase frequency is 16.

Since customer value equals to AOV\*purchase frequency and the frequency value for each store is 1, so the value of customer value is equal to the AOV value for each store.

(The summary table for AOV, average basket value, purchase frequency and customer value for every shop is located in appendix for detail)

**Question 2:** For this question you’ll need to use SQL. [Follow this link](https://www.w3schools.com/SQL/TRYSQL.ASP?FILENAME=TRYSQL_SELECT_ALL) 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.

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

SELECT

count(OrderID) as order\_shipped\_with\_Speedy\_Express

FROM [Orders]

where ShipperID = 1;

answer: 54

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

select

LastName,

count\_employee

from

(SELECT

EmployeeID,

count(OrderID) as count\_employee

from Orders

group by EmployeeID) a

join Employees b

on a.EmployeeID = b.EmployeeID

order by count\_employee desc

limit 1

;

Answer: Peacock (order amount: 40)

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

SELECT

ProductName,

total\_quantity

FROM

(select CustomerID

from Customers

where Country='Germany')a

join

[Orders] b

on a. customerID = b.customerID

join

(select orderID,

productID,

sum(quantity) as total\_quantity

from Orderdetails

group by productID) c

on b.orderID = c.orderID

join products d

on c.productID=d.productID

order by total\_quantity desc

limit 1

;

Answer: Boston Crab Meat (quantity—256)

Appendix

Table A

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| shop\_id | AOV | average\_basket\_value | purchase\_frequency | customer\_value |
| 92 | 162.8571 | 2 | 1 | 162.8571 |
| 2 | 174.3273 | 2 | 1 | 174.3273 |
| 32 | 189.9762 | 2 | 1 | 189.9762 |
| 100 | 213.675 | 2 | 1 | 213.675 |
| 53 | 214.1176 | 2 | 1 | 214.1176 |
| 93 | 214.4746 | 2 | 1 | 214.4746 |
| 7 | 218 | 2 | 1 | 218 |
| 56 | 218.1892 | 2 | 1 | 218.1892 |
| 25 | 232.9167 | 2 | 1 | 232.9167 |
| 9 | 234 | 2 | 1 | 234 |
| 34 | 234.24 | 2 | 1 | 234.24 |
| 75 | 240.7619 | 2 | 1 | 240.7619 |
| 8 | 241.0435 | 2 | 1 | 241.0435 |
| 14 | 242 | 2 | 1 | 242 |
| 48 | 242.775 | 2 | 1 | 242.775 |
| 98 | 245.3621 | 2 | 1 | 245.3621 |
| 83 | 248.7857 | 2 | 1 | 248.7857 |
| 20 | 251.5577 | 2 | 1 | 251.5577 |
| 41 | 254 | 2 | 1 | 254 |
| 68 | 254.6383 | 2 | 1 | 254.6383 |
| 36 | 254.8 | 2 | 1 | 254.8 |
| 58 | 254.9492 | 2 | 1 | 254.9492 |
| 4 | 258.5098 | 2 | 1 | 258.5098 |
| 47 | 259.1489 | 2 | 1 | 259.1489 |
| 44 | 262.1538 | 2 | 1 | 262.1538 |
| 69 | 264.1833 | 2 | 1 | 264.1833 |
| 63 | 264.9655 | 2 | 1 | 264.9655 |
| 39 | 268 | 2 | 1 | 268 |
| 31 | 268.9787 | 2 | 1 | 268.9787 |
| 45 | 269.3103 | 2 | 1 | 269.3103 |
| 16 | 270.1463 | 2 | 1 | 270.1463 |
| 64 | 272.186 | 2 | 1 | 272.186 |
| 67 | 272.6216 | 2 | 1 | 272.6216 |
| 22 | 273.75 | 2 | 1 | 273.75 |
| 54 | 276.64 | 2 | 1 | 276.64 |
| 86 | 277.5 | 2 | 1 | 277.5 |
| 49 | 279.9057 | 2 | 1 | 279.9057 |
| 77 | 280.8 | 2 | 1 | 280.8 |
| 5 | 290.3111 | 2 | 1 | 290.3111 |
| 87 | 292.2692 | 2 | 1 | 292.2692 |
| 30 | 295.0714 | 2 | 1 | 295.0714 |
| 40 | 295.1667 | 2 | 1 | 295.1667 |
| 57 | 296.7736 | 2 | 1 | 296.7736 |
| 94 | 297.7778 | 2 | 1 | 297.7778 |
| 80 | 299.6667 | 2 | 1 | 299.6667 |
| 3 | 305.25 | 2 | 1 | 305.25 |
| 74 | 306 | 2 | 1 | 306 |
| 21 | 308.6957 | 2 | 1 | 308.6957 |
| 1 | 308.8182 | 2 | 1 | 308.8182 |
| 62 | 308.8372 | 2 | 1 | 308.8372 |
| 15 | 308.9423 | 2 | 1 | 308.9423 |
| 72 | 309.5652 | 2 | 1 | 309.5652 |
| 66 | 312.8868 | 2 | 1 | 312.8868 |
| 52 | 316.9268 | 2 | 1 | 316.9268 |
| 23 | 317.6727 | 2 | 1 | 317.6727 |
| 95 | 318.7692 | 2 | 1 | 318.7692 |
| 28 | 320.3721 | 2 | 1 | 320.3721 |
| 24 | 320.7273 | 2 | 1 | 320.7273 |
| 19 | 320.9063 | 2 | 1 | 320.9063 |
| 76 | 321.0714 | 2 | 1 | 321.0714 |
| 71 | 323.0303 | 2 | 1 | 323.0303 |
| 97 | 324 | 2 | 1 | 324 |
| 91 | 325.9259 | 2 | 1 | 325.9259 |
| 55 | 327.75 | 2 | 1 | 327.75 |
| 35 | 328 | 2 | 1 | 328 |
| 79 | 328.4815 | 2 | 1 | 328.4815 |
| 85 | 329.2571 | 2 | 1 | 329.2571 |
| 96 | 330 | 2 | 1 | 330 |
| 65 | 330.8148 | 2 | 1 | 330.8148 |
| 29 | 331.6207 | 2 | 1 | 331.6207 |
| 17 | 332.0755 | 2 | 1 | 332.0755 |
| 10 | 332.3019 | 2 | 1 | 332.3019 |
| 43 | 333.9138 | 2 | 1 | 333.9138 |
| 27 | 334.8704 | 2 | 1 | 334.8704 |
| 73 | 335.6897 | 2 | 1 | 335.6897 |
| 99 | 339.4444 | 2 | 1 | 339.4444 |
| 37 | 340.2083 | 2 | 1 | 340.2083 |
| 26 | 341.2245 | 2 | 1 | 341.2245 |
| 84 | 342.3051 | 2 | 1 | 342.3051 |
| 18 | 342.5882 | 2 | 1 | 342.5882 |
| 70 | 343.0678 | 2 | 1 | 343.0678 |
| 61 | 344.44 | 2 | 1 | 344.44 |
| 13 | 345.3968 | 2 | 1 | 345.3968 |
| 46 | 347.4419 | 2 | 1 | 347.4419 |
| 82 | 349.7857 | 2 | 1 | 349.7857 |
| 60 | 350.234 | 2 | 1 | 350.234 |
| 12 | 352.6981 | 2 | 1 | 352.6981 |
| 88 | 355.52 | 2 | 1 | 355.52 |
| 11 | 356.7347 | 2 | 1 | 356.7347 |
| 59 | 358.9667 | 2 | 1 | 358.9667 |
| 51 | 361.8043 | 2 | 1 | 361.8043 |
| 33 | 376.275 | 2 | 1 | 376.275 |
| 89 | 379.1475 | 2 | 1 | 379.1475 |
| 6 | 383.5085 | 2 | 1 | 383.5085 |
| 81 | 384 | 2 | 1 | 384 |
| 38 | 390.8571 | 2 | 1 | 390.8571 |
| 90 | 403.2245 | 2 | 1 | 403.2245 |
| 50 | 403.5455 | 2 | 1 | 403.5455 |
| 78 | 49213.04 | 2 | 1 | 49213.04 |
| 42 | 235101.5 | 668 | 1 | 235101.5 |