# Winter 2022 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.

AOV (Average Order Value)- It is the average amount spent each time a customer places an order of the shoe

Since Sneakers are relatively affordable, and also that an individual cannot spend an average of $3145 on an order.

Current AOV calculation = (x1+x2+x3+…+xn)/n

This calculation of the average of order amount is wrong. Since it has some of the value that have huge numbers. Let’s take the median and standard deviation of this column.

|  |  |
| --- | --- |
| Median | 284 |
| Std Dev | 41,282.54 |
| Average/ Mean | 3145.128 |

We can see that the standard deviation is too high. It implies that the dispersion of data is too much which could be due to outliers. Now let’s visualize and see if our findings are right.

This proves that the outliers (in range of 200k-700k) is causing the average to an unrealistic number.

*Right way to Calculate:*

After removing the bulk orders >10k, the AOV=$302.58 which looks like a realistic value for the avg order amt placed by each customer. Hence this would be a better solution.

1. What metric would you report for this dataset?

- High Valued Customers (HVC): User\_id vs order amount and total items(To see who purchased - To target customers)

- High Valued shops (HVS): Shop\_id vs order amount and total items to know which shop made the most sales etc

1. What is its value?

*HVC:*

- User ID 607 is making bulk purchases of items and gives the max revenue (they could be a wholesale retailer or small business).

- Followed by are user ids 878, 834 who are elite buyers purchasing expensive sneakers.

These customers could be targeted for promotional offers, to encourage their purchasing capabilities and build more trust with the company

*HVS:*

- Shop IDs 42 is doing good with a revenue of 11.9M and total item sale of 34K shoes

- Followed by are user ids 78 which sells expensive sneakers for a high revenue of 2M

These shops can be improved with more investments and seasonal offers

**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 Total\_Orders

FROM Orders as O

inner join Shippers as S

on S.ShipperID=O.ShipperID

where S.ShipperName='Speedy Express'

ANS: 54

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

SELECT E.LastName, count(OrderID) as Most\_Orders

FROM Employees as E

inner join Orders as O

on E.EmployeeID=O.EmployeeID

group by E.LastName

order by count(OrderID) desc

|  |  |
| --- | --- |
| **LastName** | **Most\_Orders** |
| Peacock | 40 |

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

SELECT P.ProductName, sum(OD.Quantity) as Most\_Ordered

FROM Customers as C, Orders as O, OrderDetails as OD, Products as P

where C.Country='Germany'

and C.CustomerID=O.CustomerID

and O.OrderID=OD.OrderID

and OD.ProductID=P.ProductID

group by P.ProductName

order by sum(OD.Quantity) desc

Boston Crab Meat