# Answers to Winter 2022 DS Intern Challenge

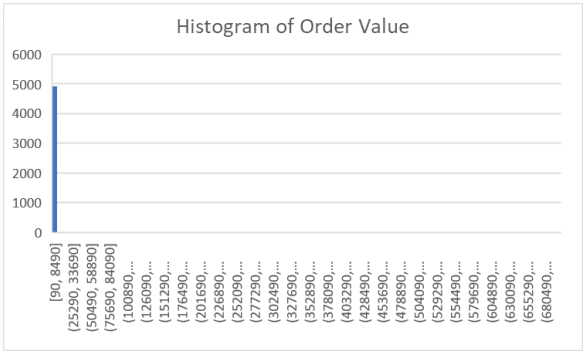
By Can Zhang, can7@ualberta.ca

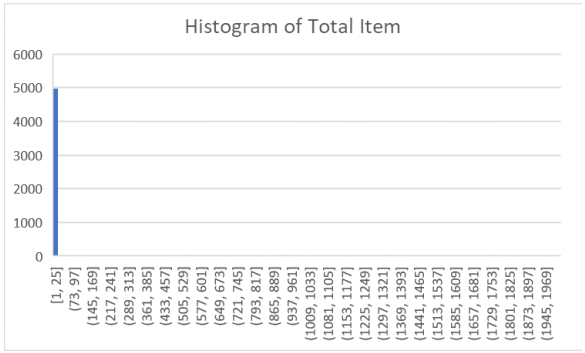
**Question 1:** Given some sample data, write a program to answer the following: click here to access the required data set

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.

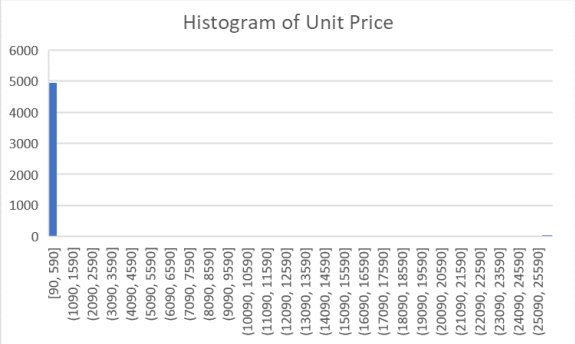
a. ***Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.***

Look at the histogram below for order value and total item.




For the histogram of order value, there are orders with extremely high value around 700000. However, major proportion of the population is distributed under 8490. This is very abnormal.
For the histogram of Total Item, there are orders with item number over 1000. However, major proportion of the population is distributed in the [1,25] range.
We continue to see the histogram of unit price for each order.




The same trend appears, most orders have unit price under 600, while there are extremely high values over 25000.

Considering the high order value and unit price, I investigated the table and find **the unit prices for all orders in store with shop_id 78 are 25725**. Since each shop sells one type of affordable shoes. It may be caused by currency error or other system/manual errors.

Considering the high item number, I sorted the order by item number. **And I found user with user_id 607 created 17 orders at shop with shop_id 42 with item numbers all at 2000**. Shop 42 may be a wholesale store and user 607 may run a shoe business.

It is also high possibility of money-laundering for both shops (78 and 42), which will not be discussed here.

**In conclusion, there are extremely high order values. The order value is skewed distributed, which means AOV is highly affected by extreme values. Thus, AOV can not represent the major proportion of the dataset.**

### *b. What metric would you report for this dataset?*

I will use **median of order value (MOV)** to report for this dataset. Unlike the mean, the median value doesn't depend on all the values in the dataset. It will ignore the impact of extreme values. Since we have a skewed distribution, median is a better choice than average.

### *c. What is its value?*

**The median of order value is 284.**

**Question 2:** For this question you'll need to use SQL. Follow this link 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.

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

**SQL code:**

```
SELECT COUNT(*) as Total_Order_By_Speed_Express
FROM Shippers s JOIN Orders o
ON s.ShipperID = o.ShipperID
WHERE s.ShipperName = "Speedy Express";
```

**54 orders were shipped by Speed Express in Total.**

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

**SQL code:**

```
SELECT e.LastName AS Last_Name_Employee_Top_Seller
FROM Employees e LEFT JOIN Orders o
ON e.EmployeeID = o.EmployeeID
GROUP BY e.EmployeeID
ORDER BY COUNT(*) DESC
LIMIT 1;
```

**The last name of the employee with the most orders is "Peacock".**

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

**SQL code:**

```
SELECT p.ProductName AS TOP_SALE_IN_GERMANY
FROM OrderDetails o1 JOIN Orders o2 ON o1.OrderID = o2.OrderID
JOIN Products p ON o1.ProductID = p.ProductID
JOIN Customers c ON o2.CustomerID = c.CustomerID
WHERE c.Country = "Germany"
GROUP BY o1.ProductID
ORDER BY SUM(o1.Quantity) DESC
LIMIT 1;
```

**"Boston Crab Meat" was ordered the most (with highest total quantity) by customers in Germany.**