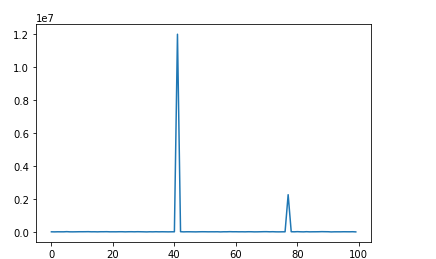
Winter 2021 Data Science Intern Challenge

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!

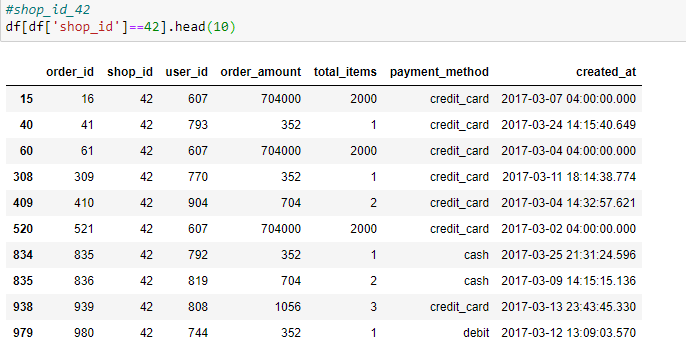
Question 1: Given some sample data, write a program to answer the following: click here to access the required data set

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.

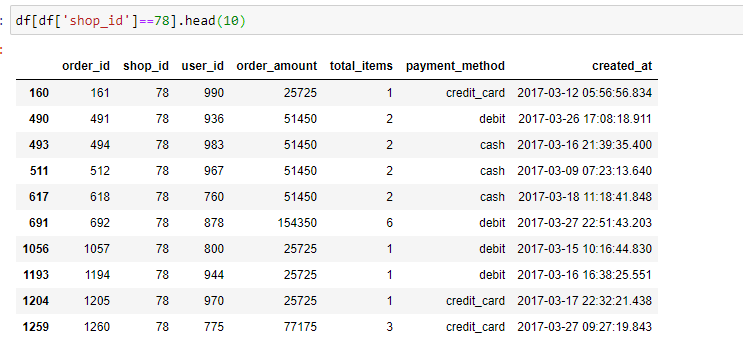
1. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.



This is the graph for Order amount and shopID, from that we can clearly see two spikes one at 42 and another at 78. The unconvincing AOV price should come from these two unusual shops.

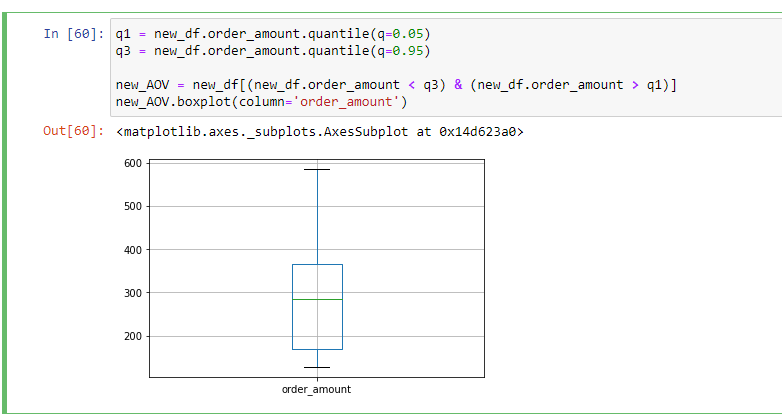
After looking into these two shops: 

We can found the price for one pair is not high but there are some 2000 pairs orders in shop 42 everyday at 4. It seems like the transactions are probably some sort of supplier purchasing many shoes at once since the order amount is consistently 2000.



For shop 78, it’s high total order amount comes from the high price of each shoes pair (25725 for each). It’s possible that shop 78 sells a kind of luxury sneakers which has higher order amount than usual ones.

Therefore, to get a more reliable AOV, I would firstly remove orders from these two shops as they are outliers. Then I used boxplot to find the median of rest data in range (5%-95%)

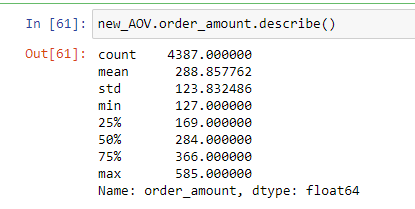


1. What metric would you report for this dataset?

I would use the median values of the data in the new boxplot which remove the first and last 5% of data. After removing all outliers, the median can better represent the AOV price of most shoes. Also, the std decreased a lot after removing outliers which also tells that this is a better approach to calculate AOV.

1. What is its value?

As the result in he table below, the data I got is 284 which is a more convincing number.



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.

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

The output shows the number of orders is **54**.

SELECT COUNT(\*) AS SpeedyExpressNum

FROM Orders

JOIN Shippers

ON Shippers.ShipperID = Orders.ShipperID

WHERE Shippers.ShipperName = 'Speedy Express'

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

The employee with the last name **Peacock** had the most orders at **40**

SELECT LastName,most\_freq\_order AS NumberOfOrders

FROM (

(SELECT EmployeeID, MAX(sales) AS most\_freq\_order

FROM (

SELECT EmployeeID, Count(OrderID) AS sales

FROM Orders

GROUP BY EmployeeID

)

) AS t1

JOIN

(SELECT EmployeeID, LastName

FROM Employees

) AS t2

ON t1.EmployeeID == t2.EmployeeID

)

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

**Boston Crab Meat** has the most orders at **160 total orders**.

SELECT ProductName,Most\_Freq.most\_freq\_order As TotalOrder

FROM Products

JOIN

(SELECT ProductID, MAX(TotalQantity) AS most\_freq\_order

FROM

(SELECT ProductID,SUM(Quantity) as TotalQantity

FROM OrderDetails

JOIN

(SELECT Orders.OrderID, Customers.Country

FROM Orders

JOIN Customers

ON Customers.CustomerID = Orders.CustomerID

WHERE Customers.Country = 'Germany') AS t1

ON OrderDetails.OrderID == t1.OrderID

GROUP BY ProductID

)) AS Most\_Freq

ON Products.ProductID = Most\_Freq.ProductID