**Fall 2022 Data Science Intern Challenge**

Solution –

1. For getting the accurate result we need to preprocess the data and do some data cleaning part. We have observed that the values of $3145.13 is the mean price of the show which is wrong to consider as does not tell the AOV(Average order value and on top of it the data is biased and have some anomalies. The better approach would be dividing the data set into the average price per sneeker i.e order amount/ number of items. We have added one column in Jupyter which states the average value per shoes which comes out to be $387 which is comparatively lesser but we have to pre process the data and remove anomalies.

We also noticed that the data has many outliers present in amount, price columns. First, we checked the outliers by box plot of every individual column and remove them. We noticed the amount of 70000 is highly affecting the overall price and it would be logical to remove it.

We also noticed that the average price value is also high for some store id and we also removed them to get the final dataset.

1. The correct metric would be calculating the z score and checked the outliers and remove them. We segregated the quantiles and checked the 50%, 60%, 70%, 90%, 95% and 99% values and understand whether to remove it or not.
2. The final value comes out to be $154.

Question2:

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

SELECT ShipperName, COUNT(ShipperName) OrdersShipped FROM(

SELECT\*

FROM Orders o

LEFT JOIN Shippers s

ON s.ShipperID=o.ShipperID)

GROUP BY ShipperName

HAVING ShipperName'Speedy Express'

Answer:Speedy Express has shipped 54 orders

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

SELECT\*FROM(

(SELECT EmployeelD,COUNT(\*)NumOrders FROM[Orders]

GROUP BY EmployeeID) e

LEFT JOIN(SELECT EmployeeID,LastName FROM Employees)a

ON e.EmployeeID=a.EmployeeID)

ORDER BY NumOrders DESC

LIMIT1

Answer:Peacock is the last name of the employee with the most orders,with 40 orders

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

Note:This assumes that we sum quantity of product,not number of purchases

SELECT ProductName,Country,TotalQuantity FROM(

SELECT Productld,Country,SUM(Quantity)Total Quantity FROM OrderDetails od

LEFT JOIN(

SELECT o.OrderID,o.CustomerID,c.Country FROM Orders o

LEFT JOIN(SELECT Customerld,Country FROM Customers) c

ON o.CustomerID=c.CustomerID)custom

ON od.OrderID=custom.OrderID

GROUP BY Country,ProductID

HAVING Country='Germany')germany

LEFT JOIN(SELECT Productld,ProductName FROM Products)p

ON germany.Productld=p.Productld

ORDER BY Total Quantity DESC

LIMIT1

Answer:Boston Crab Meat is the product that was ordered most by customers in

Germany,with 160 ordered.