Summer 2022 Data Science Intern Challenge

Question 1: (Using R)

data <- read.xlsx("D:/2019 Winter Data Science Intern Challenge Data Set")

dataordered <- data[order(data$order_amount),]

a)

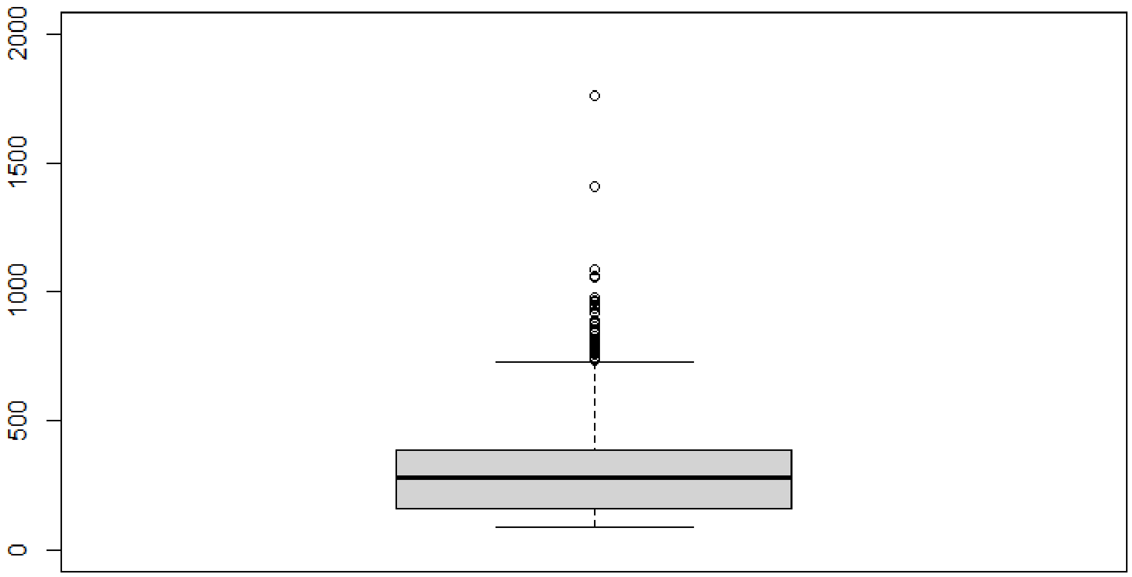
Mean Value:

> mean(dataordered$order_amount)

[1] 3145.128

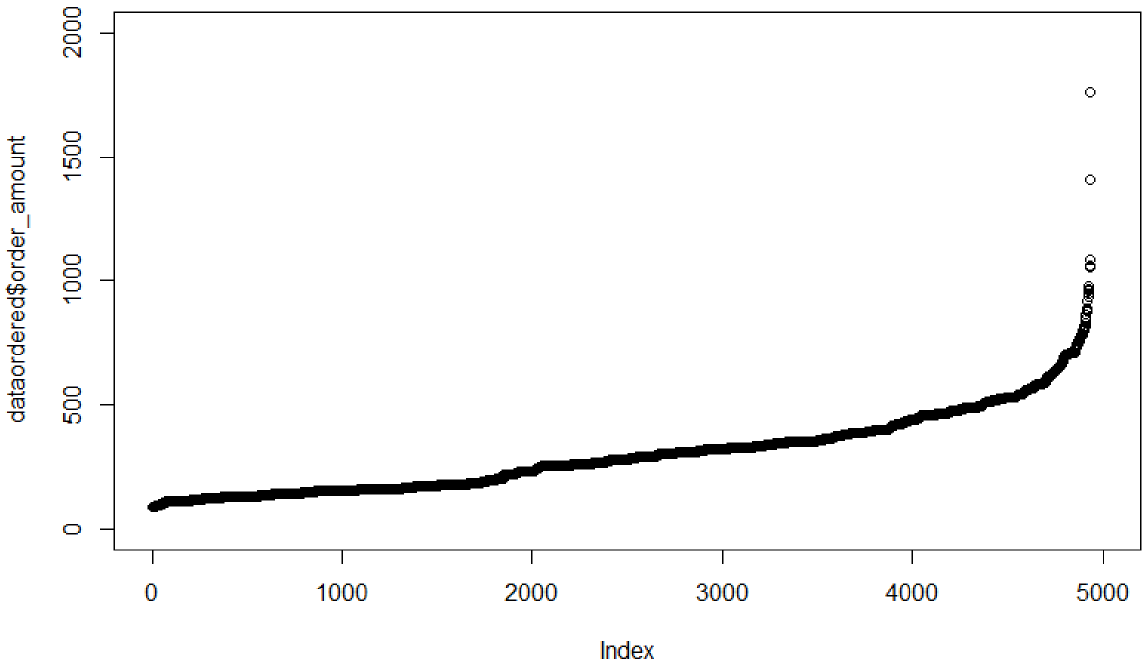
In the dataset we have a few outliers, which is affecting the AOV.

boxplot(dataordered$order_amount,ylim = c(0,2000))



Most of the value data falls under 1000.

plot(dataordered$order_amount,ylim = c(0,2000))



Of the 5000 records, 4937 data points are less than 2000 giving, 98% of data in the range of 1000’s.

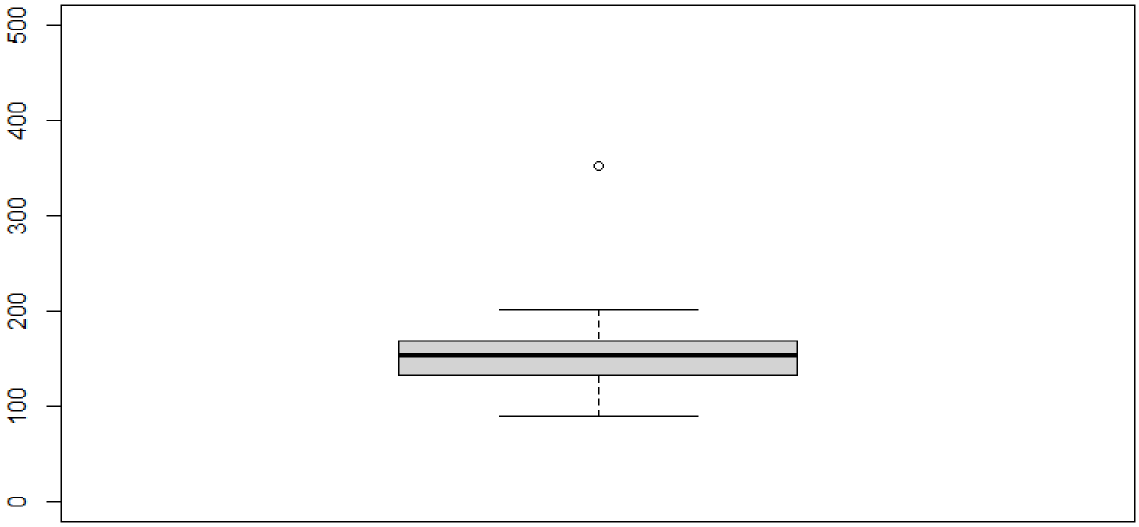
b)

We can clean the dataset by first taking the average amount for each shop

dataordered <- transform(dataordered, avgofeachshop = (order_amount/total_items))

boxplot of the avgofeachshop:

boxplot(dataordered$avgofeachshop,ylim = c(0,500))



dataordered_2 <- dataordered[order(dataordered$avgofeachshop),]

On looking at this one outlier we can find that 46 shops are giving an average of $25725, for a number of 1,2,3 shoes. This can arise from a typo or change in unit(cents instead of dollars)

Correcting the value to $257.25 instead.

dataordered_2$corrected <- ifelse(dataordered_2$avgofeachshop>=500, 257.25, dataordered_2$avgofeachshop)

mean of the resulting dataset gives:

mean(dataordered_2$corrected)

= 153.4395

c)

Mean of cleaned dataset: 153.4395

Question 2:

Sql queries:

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

SELECT Count(OrderID) FROM Orders o

Inner JOIN Shippers s ON s.ShipperID = o.ShipperID

WHERE s.ShipperName = 'Speedy Express'

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

```
SELECT  Employees.LastName FROM Orders

INNER JOIN Employees  on Employees.EmployeeID = Orders.EmployeeID

GROUP BY Employees.LastName

Having Count(OrderID) >= ALL(SELECT Count(OrderID)FROM Orders GROUP BY EmployeeID);
```

c)

What product was ordered the most by customers in Germany?

```
Select [Products].ProductName,sum(OrderDetails.Quantity)

from Orders

Inner Join Customers on Customers.CustomerId = Orders.CustomerId

Inner Join [OrderDetails] on Orders.OrderId = [OrderDetails].OrderId

Inner Join Products on Products.ProductId = [OrderDetails].ProductId

Where Customers.Country = 'Germany'

Group By [Products].ProductName

Having sum(OrderDetails.Quantity) >= ALL(SELECT sum(OrderDetails.Quantity) FROM Orders

Inner Join Customers on Customers.CustomerId = Orders.CustomerId

Inner Join [OrderDetails] on Orders.OrderId = [OrderDetails].OrderId

Where Customers.Country = 'Germany'

GROUP BY [OrderDetails].ProductId);
```