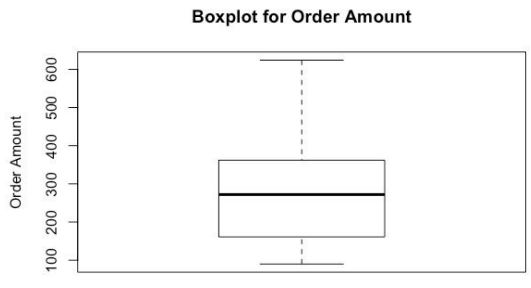
Question 1:

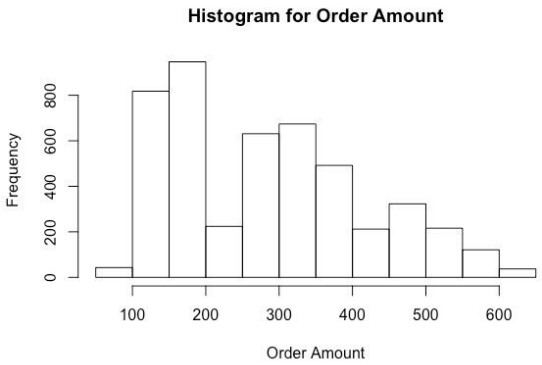
Currently Shopify is calculating the average order value by taking the average of all the orders (i.e. AVERAGE(D2:D5001) for the provided spreadsheet). However it is evident that something is wrong with this calculation as an order of relatively affordable items should not have an average of $3000+ per order.

a. It's important to note that we are trying to determine the value of an average order and not *every* order. This means that we should probably eliminate the outliers from our data before calculating the average. A standard way to do this is using the IQR method, in which we only look at the values that are within the range of the median plus or minus 1.5 times the IQR. The code for this would look something like the following (using R):

```
#Import dataset X2019_Winter_Data_Science_Intern_Challenge_Data_Set using RStudio
dataset <- `X2019_Winter_Data_Science_Intern_Challenge_Data_Set`

#Analyze Data by Order Amount
average_order_value <- function(data){
  #eliminate outliers using IQR
  order_amount_iqr <- IQR(data$order_amount)
  median <- median(data$order_amount)
  upperbound <- median + 1.5*order_amount_iqr
  lowerbound <- median - 1.5*order_amount_iqr
  new_order_data <- data[data$order_amount >= lowerbound & data$order_amount <= upperbound,
]

  #boxplot & histogram visual representation
  boxplot(new_order_data$order_amount, main="Boxplot for Order Amount", ylab="Order Amount")
  hist(new_order_data$order_amount, main="Histogram for Order Amount", xlab="Order Amount")

  #find average order value
  mean(new_order_data$order_amount)
}

#aov value for the given dataset
average_order_value(dataset)
```

Note that it is not necessary to include the boxplot and histogram in the code above. It is only there to better analyze/understand the data. For instance, for the given data we get the following boxplot and histogram:





Clearly from the adjusted graphs above we can see that the majority of orders fall below $400, and that the frequency of order values are a combination of normal distributions. That is to say that, there are different ranges where the data peaks and falls. For instance, between $100-$200, $200-$400, and $400-$600. This may be due to the different buyer price range preferences, in this case it could be that more customers have a budget of around $200, so that may be why there is a higher frequency of orders within the $100-$200 range. This would also be a good explanation for why there are different ranges, since each range represents a different cohort/group of buyers with different price points. Of course this is only speculation, and would probably require further research to back-up.

b. As for the metric, we can still use the same one as before, that is, the average amount/value per order. However unlike before, it is being calculated after filtering out the outliers. There are also different ways to filter outliers, including using the standard deviation or even a set range that marketing or sales has noticed through trends.

c. The new average order value is $283.81 after running and rounding the code above.

It is clear that this value fits the data a lot better as significant outliers have been removed before finding the average. This eliminated certain extreme orders which skewed Shopify's original analysis.

Notice however that there are many other things that Shopify may want to consider when analyzing the data. A simple number/value does not provide much insight on the average values per order for customers. For instance, as we described above, the budgets of each customer should be considered if we are trying to determine which price of shoe sells better. Or perhaps the orders depend on the sneaker shop, as certain shoes may be more popular compared to others and thus have more people ordering them. With the data provided it may also be a good idea to look at what time of month the orders are being made, since more orders might be made at certain times of the month (perhaps due to pay days). Although this may not indicate the average order value in terms of revenue, it can affect other aspects of the business. For instance the amount of money spent on keeping things in inventory. Essentially, it is important to consider multiple aspects of provided data to spot trends that may potentially be helpful for the business. (Depending on the resources, time, and goals of the study).

However in this situation I have only provided one answer that is a relatively basic analysis of the data provided, (as there isn't much specification as to what the company wants to understand in terms of the average order value).

Question 2:

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

```
SELECT COUNT(*)
FROM Orders
INNER JOIN Shippers
ON Orders.ShipperID = Shippers.ShipperID
WHERE Shippers.ShipperName='Speedy Express';
```

Answer: 54 orders were shipped by Speedy Express in total.

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

```
SELECT LastName
FROM(
SELECT LastName, MAX(NumberOfOrders) AS MostOrders
FROM (
SELECT Employees.LastName,COUNT(Orders.OrderID) AS
NumberOfOrders FROM Orders
INNER JOIN Employees ON Orders.EmployeeID =
Employees.EmployeeID
GROUP BY Employees.EmployeeID
));
```

Answer: The last name of the employee with the most orders is Peacock.

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

```
SELECT ProductName
FROM(
SELECT ProductName, MAX(TotalOrders)
FROM (
SELECT Products.ProductName, SUM(OrderDetails.Quantity) AS
TotalOrders
FROM ((( Orders
INNER JOIN Customers ON Orders.CustomerID =
Customers.CustomerID)
INNER JOIN OrderDetails ON Orders.OrderID =
OrderDetails.OrderID)
INNER JOIN Products ON OrderDetails.ProductID =
Products.ProductID)
WHERE Customers.Country = "Germany"
GROUP BY Products.ProductName
));
```

Answer: Boston Crab Meat was ordered the most by customers in Germany.