Shopify Technical Challenge Submission (link)

Ryan Elliott

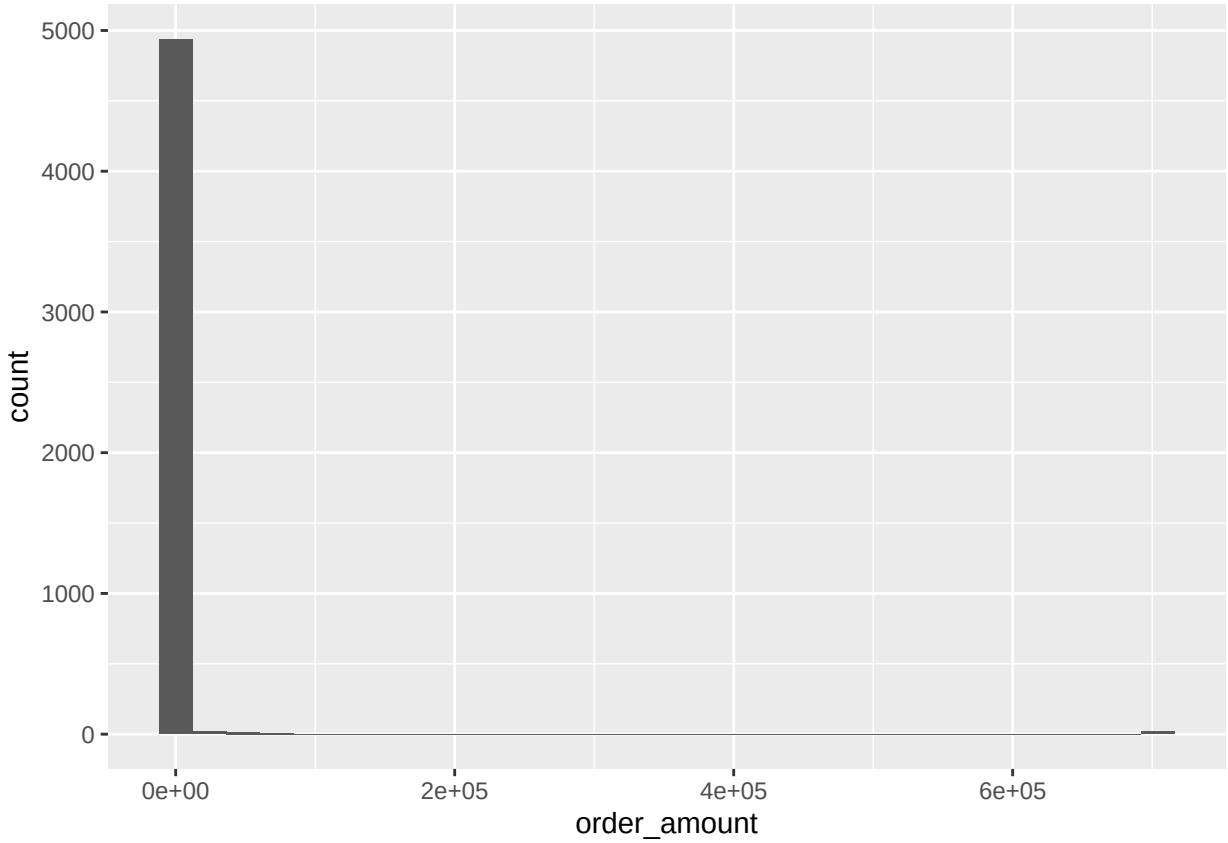
06/05/2021

Internship Challenge Question 1

Part (a) The average order value (AOV) has be identified as $3145.13. The has been determined by taking the average of the order_amount field from the data. Most applications for taking the average of a group of data points assumes its distribution to be Gaussian (or normal). In this data set we have a handful of orders that are several orders of magnitude greater than even the average. These outliers can be visualized in the histogram below.

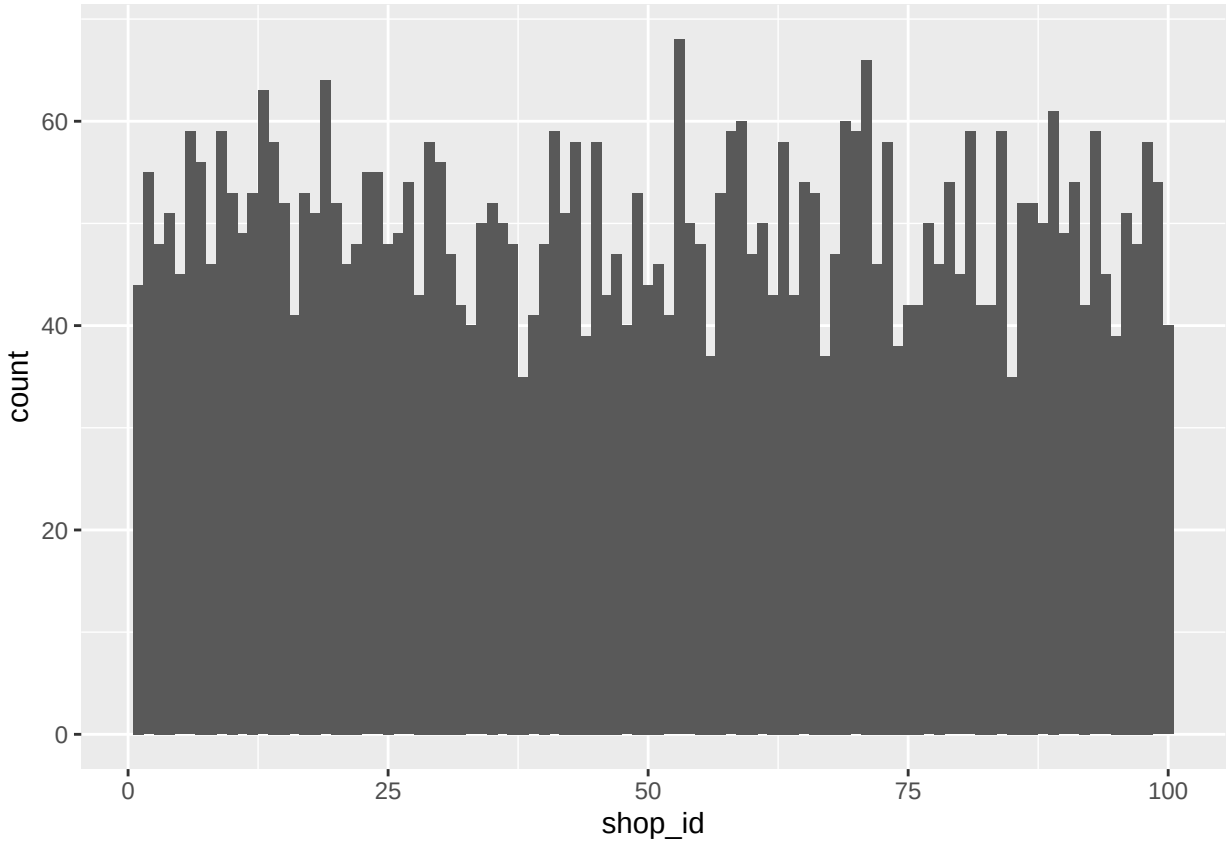
```
d%>%
ggplot(aes(order_amount))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Another factor that could be influencing the high average value is the distribution of sales at shops. If there is one shop that sells a lot of shoes at a high value then this will increase the average. It was determined that the sales amongst the shops was uniformly distributed (based on inspection of the histogram below) and that the only contributory factor to the large average order value is the outliers in the data.

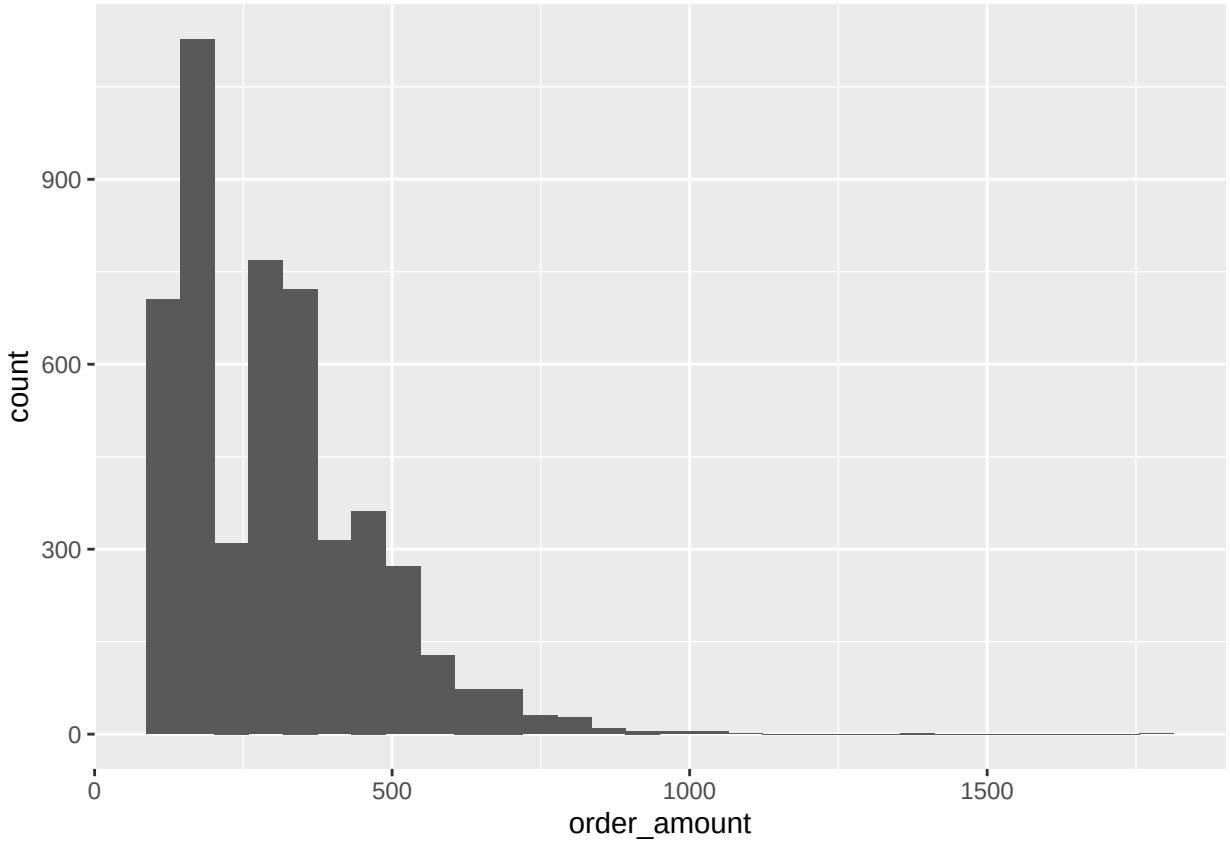
```
d%>%
ggplot(aes(shop_id))+geom_histogram(binwidth = 1)
```



Part (b) If the task is to evaluate the AOV of sales from the past 30 days we will have weight the amount of each order by the total number of items ordered. By doing so we are making an assumption that each item within an order is identically priced. The median is a more appropriate metric for identifying the value of the most frequent order. The median of the order_amount is $284. That means that half of all transactions is greater than $284 and while the other half is less than $284. Another solution to better understand the distribution of data is to remove the outliers. One method to evaluate the outliers is to use the Zscore. Zscore evaluates each data point and determines the number of standard deviations away from the mean. While the mean is suspect given, evaluation of zscore provides a statistical argument for omitting outliers.

```
d[d$order_amount<4000,]%>%
ggplot(aes(order_amount))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Part (c) The AOV of the unprocessed data is $3145 which is well beyond 99% of the order amount of each record. I will therefore omit all data that has a zscore >0. This means that all records whose value is less than the mean will be evaluated for the AOV calculation.

```
d<-d%>%
  mutate(zscore = (order_amount - mean(order_amount))/sd(order_amount) )
print(paste("The number of records whose value is less than the mean is",sum(d$order_amount<3145)/5000*
```

```
## [1] "The number of records whose value is less than the mean is 98.74 %"
```

```
print(paste("The adjusted AOV is",mean(d$order_amount[d$zscore<0])))
```

```
## [1] "The adjusted AOV is 302.580514482479"
```

As displayed, the adjusted AOV is $303 which is close to the median original median of $284.

Internship Challenge Question 2

1. Speedy Express shipped 54 orders.

```
SELECT COUNT(OrderID) as Shipments FROM Orders where ShipperID = 1;
```

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

```
select count(Employees.LastName),Employees.LastName from Orders
INNER JOIN Employees
  ON Orders.EmployeeID=Employees.EmployeeID
group by LastName
order by count(LastName) DESC;
```

3. The product most ordered from Germany was Gorgonzola Telino.

```
select Count(Products.ProductID), Products.ProductName
from ((Customers
    inner join Orders
        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
    Order By Count(Products.ProductID) DESC;
```