Shopify 2022 Winter Challenge

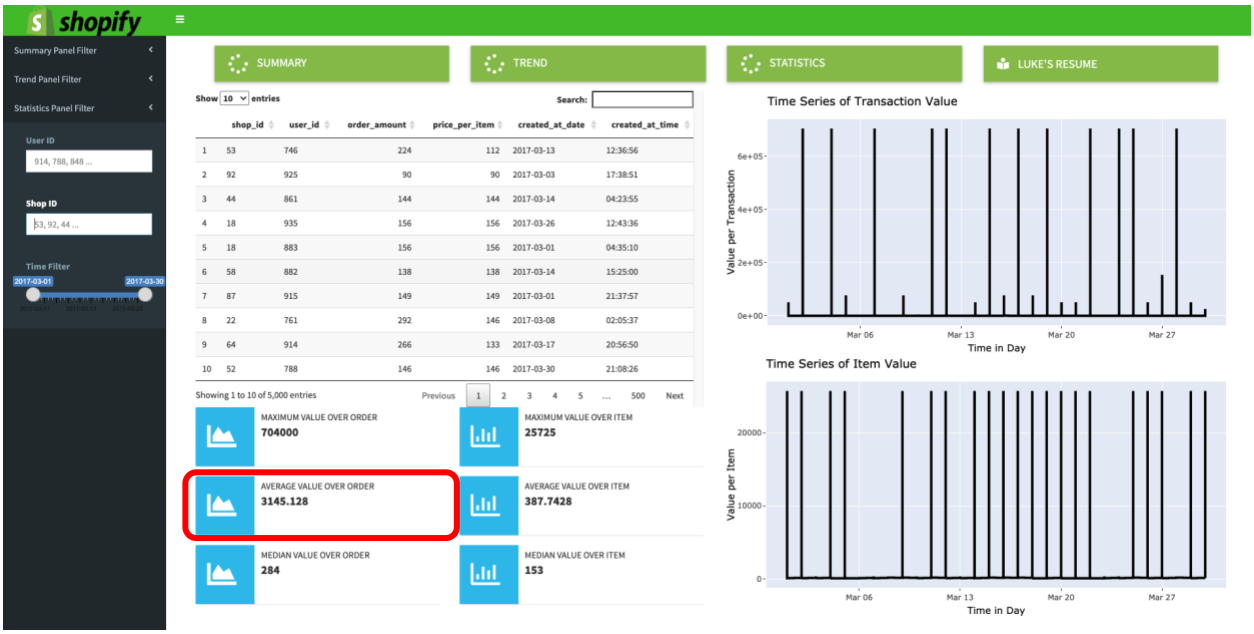
The Link for Outlier Detection Dashboard:
https://superp0tat0.shinyapps.io/shopify_shiny/

Hi, there

I think the report is only for one time usage and kind of boring to write, so instead I built an interactive visualization dashboard to make it super easy to detect the abnormality of the dataset for everyone, even the people who do not familiar with data before. So below I will use the visualization to answer the questions from the challenge.

First Question: Where is the $3145.13 AOV come from?

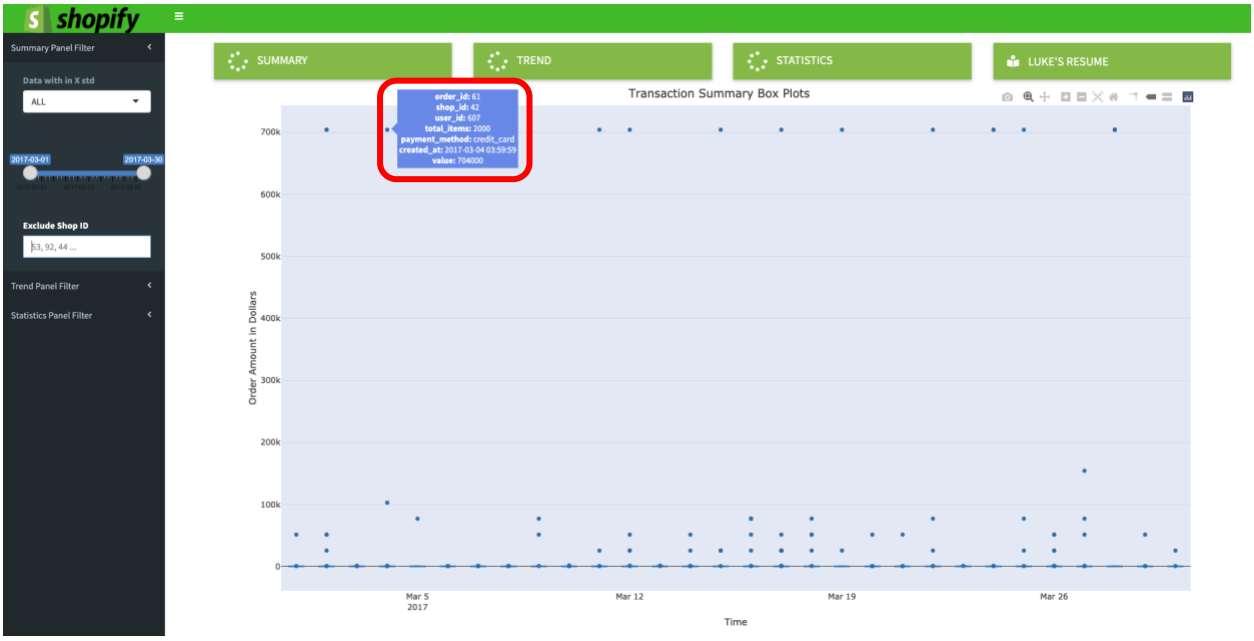
Answer: From the STATISTICS Panel, we could see It simply calculate the average of order value among all the transactions.

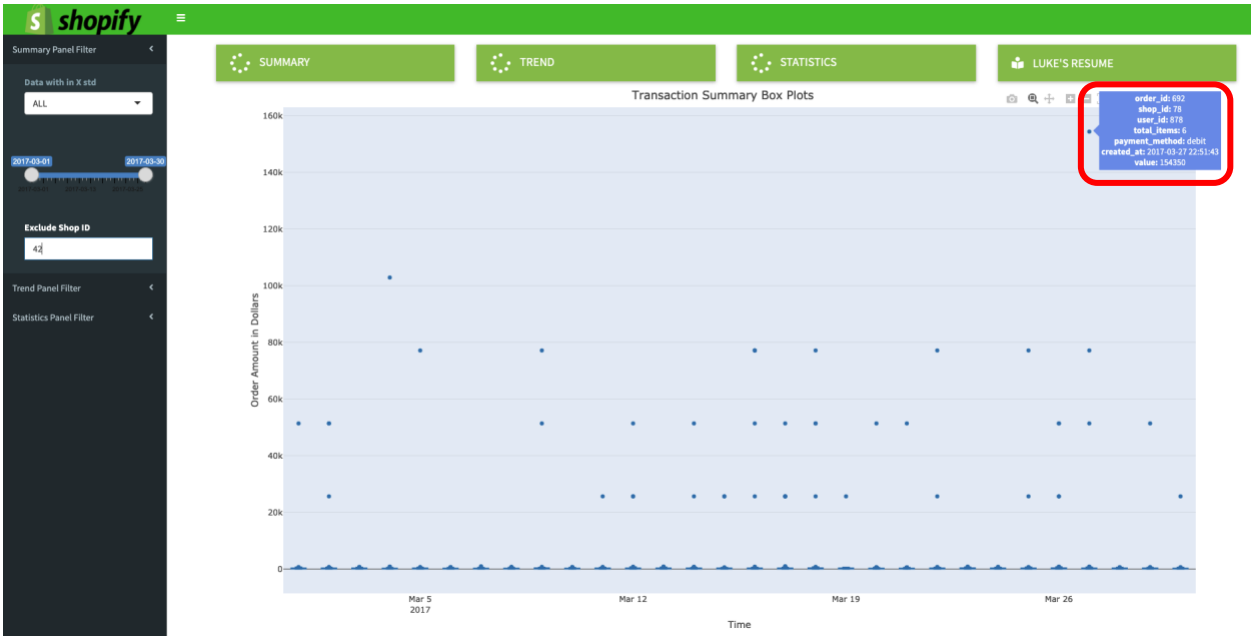


Second Question: What is going wrong with this Calculation?

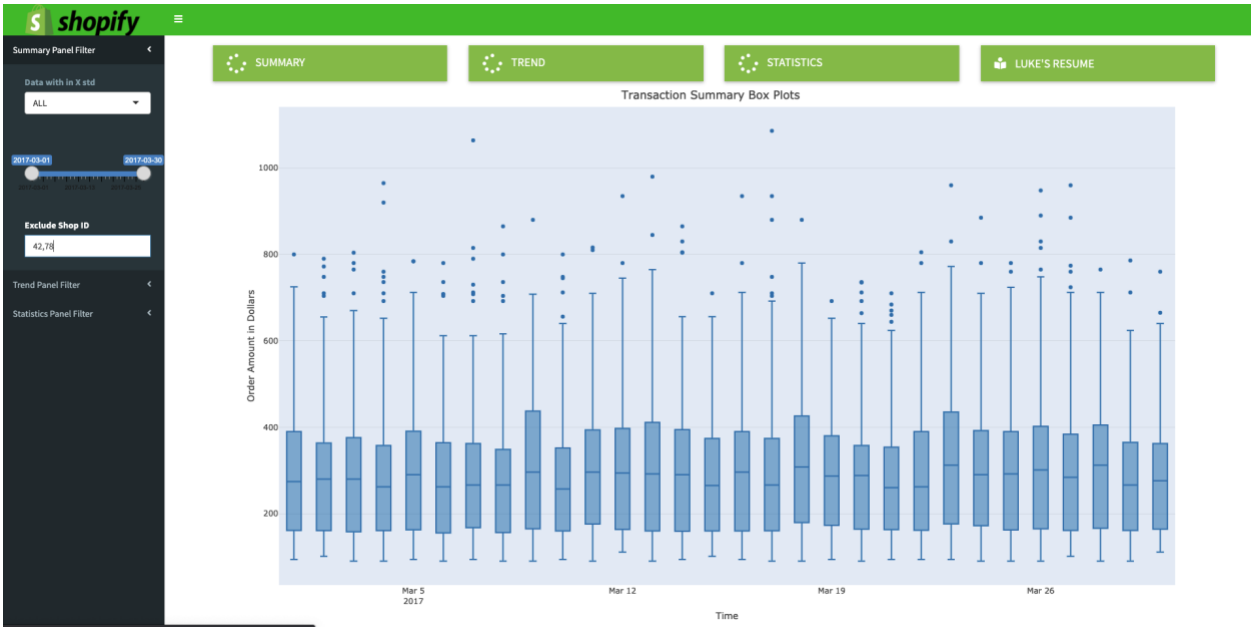
Answer: Because there so many outliers from the dataset. For example, switch to SUMMARY panel, we see shop with id 42 and 78 are the two most significant outliers; And they clearly have patterns. Just like they have the save items per

transaction for multiple times. For shop 42 it also has the same payment methods etc.
(You might want to play with the visualization to see more details)



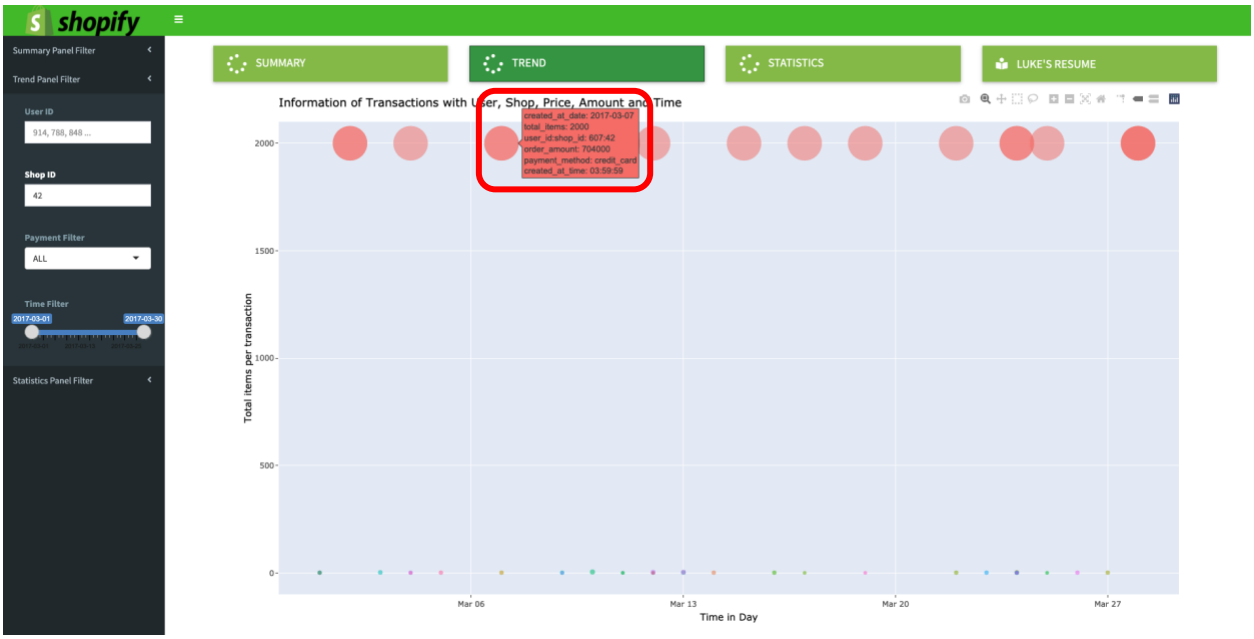


After we removed shop with id 42 and 78. We could see the distribution tends to get normal with some small outliers, but there are no longer weird patterns and mistakes.



Next, if you are interested in shop 78 or shop 42 just like me. You could go to TREND panel to further analyze the trend for each shop.

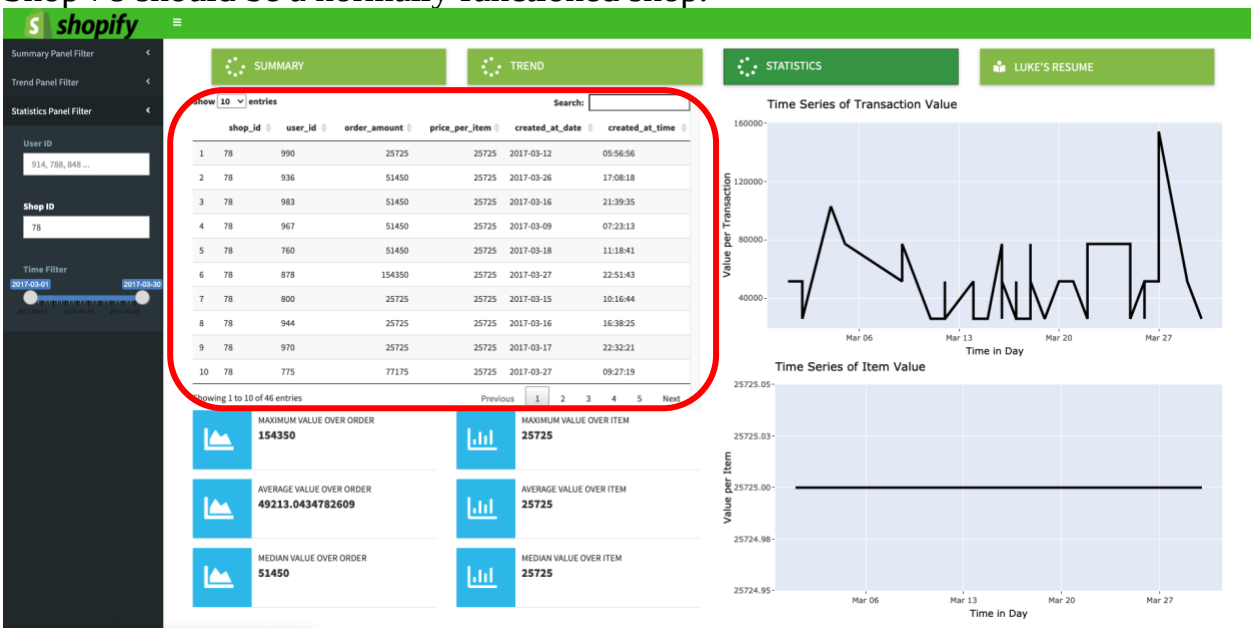
For shop 42, the data shows some of the transactions (the red one) are made by the same user (607) at shop 42. With total items to be 2000, value at 704000, payed using credit card and made at 03:59:59 AM. So this would definitely be some database or crawlers errors.



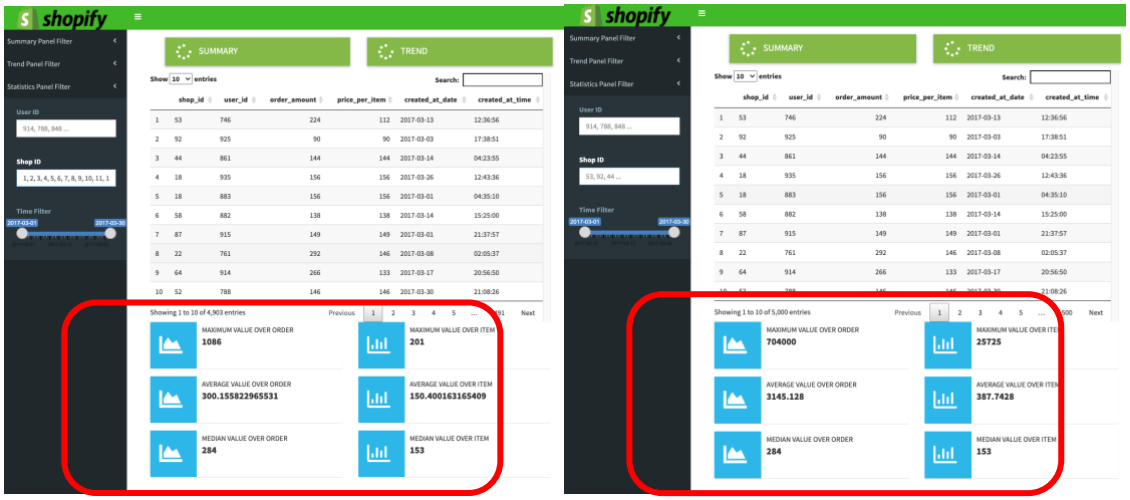
To further confirm our hypothesis, we should take a look of user 607. And yep, from the visualization we see user 607 only made purchase at shop 42. So definitely a database or crawler error.

How about shop 78? Well, from the visualization we could see there are no strange patterns. Seems they charge $25725 for each pair of sneakers. Even though it is quite a lot, but the visualization tells us there are different customers, bought different number of items at different date and time. So, I would say they are influential points, but not outliers.

We could further confirm from the STATISTICS Panel, even though the price is high. But consider the consumption volume is much lower than the other shops. Shop 78 should be a normally functioned shop.



Question: What metrics would I report and what is it value?
Answer:
Let's see why the AOV is a bad metric.
After remove the outliers shop 78 and 42 (The left image), compared to the statistics of all datasets (The right image). We could see AOV changes significantly.
But, for some metrics it does not change that much. Which means they are robust. Like the median value over order, or the median value over item. Both are good. The reason is the outliers do not take huge proportion of the dataset so they would not affect the median too much.



Alright, that would be my answer to the challenges. As you can see, I have done all the dirty works and wrap it with a fancy visualization that even the people who do not know data at all could do the analysis.

Last but not the least, please check my resume at Luke's Resume Panel.

SQL Questions and Answers

How many orders were shipped by Speedy Express in total?

Results: 54

Query:

```
SELECT count(*) as count
FROM Shippers as s JOIN Orders as o
ON s.ShipperID = o.ShipperID
WHERE s.ShipperName = "Speedy Express"
```

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

Results: Peacock

Query:

```
SELECT count(e.EmployeeID) as count, e.EmployeeID as ID, e.LastName
FROM Employees as e JOIN Orders as o
ON e.EmployeeID = o.EmployeeID
GROUP BY e.EmployeeID
ORDER BY count(e.EmployeeID) DESC
LIMIT 1
```

What product was ordered the most by customers in Germany?

Results: Boston Crab Meat

Query:

```
SELECT p.ProductName as name, SUM(od.Quantity) as quantity
FROM Customers as c JOIN Orders as o JOIN OrderDetails as od JOIN Products as p
ON c.CustomerID = o.CustomerID and o.OrderID = od.OrderID
and od.ProductID = p.ProductID
WHERE c.Country = "Germany"
GROUP BY p.ProductName
ORDER BY SUM(od.Quantity) DESC
LIMIT 1
```