Shopify Data Science Internship Challenge Question 1

Given some sample data, write a program to answer the following: click here to access the required data set

On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of $3145.13. Given that we know these shops are selling sneakers, a relatively affordable item, something seems wrong with our analysis.

1. **Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.**

   The AOV calculation was done by simply taking the mean of the order amounts. Order amounts can be mathematically written as:

   $$Order\ Amount = Total\ Number\ of\ Items\ \ *\ \ Amount\ of\ each\ Item$$

   The dataset has outliers in both $Total\ Number\ of\ Items$ (there are 17 orders with 2000 items in each order, which is naturally not normal) and $Amount\ of\ each\ Item$ (46 orders had one pair of sneakers being sold at 25,725$).

   Removing these outliers would remove the heavy skew in the calculation and bring the AOV metric to an intuitive scale. After doing just that, the numbers obtained were -

   - Corrected Average Order Value - 300.16$
   - Average Total Items purchased on each order - 2
   - Average price of each pair of sneakers - 150.4$

   Collectively, these numbers say that on average customers bought 2 pairs of sneakers, which on average costed close to 150$. Hence on each order these customers spent about 300$.

2. **What metric would you report for this dataset?**

   I feel it is more appropriate to analyse the data with more than one metric. To do this, I am keep track of the following 4 quantities for each store in the dataset:

   - `footfall`: The number of users that order from each store.
   - `cart_size`: The average number of items in a single order from each store.
   - `item_cost`: The average cost of each item in the store.
   - `revenue`: The total revenue generated by each store.

   Out of these, `revenue` is what most companies would be interested in, as that is what drives the business. The other three metrics are used to understand what drives the `revenue` generated from each store, and they provide hints as to what can be done to improve the `revenue` of a particular store.

   Note: The `footfall` is not measured using distinct users. This is because the fact that there could be many orders placed by a single user suggests that the given store is popular, and I would like to capture that information.

3. **What is its value?**

   Let us look at the minimum and maximum revenues generated by any store, and what are the values for the other metrics in these cases.

   Maximum revenue store

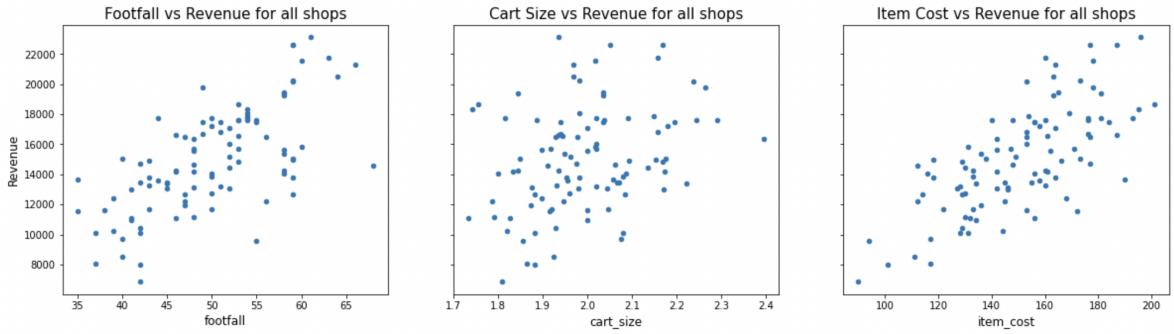
shop_id	footfall	cart_size	item_cost	revenue
86	61.0	1.934426	196.0	23128.0

Minimum revenue store

shop_id	footfall	cart_size	item_cost	revenue
89	42.0	1.809524	90.0	6840.0

Inference: These numbers clearly indicate that the store generating the highest revenue had on average more expensive sneakers, and more number of items being bought per order. Additionally, the store also had greater number of visitors.

The plots below show how the metrics `footfall`, `cart_size`, and `item_cost` influence `revenue` of a store, and what sort of inferences we can draw from them.



The **left-most graph** indicates that the revenue of each store generally increases with the number of customers visiting the store. However, there are quite a few cases where stores with lower footfall outperformed stores with higher footfall. This could be because some stores sell expensive sneakers, and hence need lower number of customers to generate same or greater revenue. There could be many other reasons that should ideally be analyzed using datasets about the stores itself.

The **middle graph** indicates that similar to footfall, the revenue also generally increases with the average number of items in each order at a store. However, there are more deviations here. This makes sense, as stores with expensive sneakers

can easy outperform other stores by even selling less pairs of sneakers in each order.

Lastly, the **right-most graph** is similar in its story that if the average price of a sneaker being sold at a store is high, the revenue generated will also be high. There are deviations in this case as well, and we should use the data of individual stores to judge what they are doing right and what they are doing wrong, which will enable us to propose steps for the business to improve revenue across all under-performers.