January 13, 2022

August Delbert

Graduate Student in Computational Mathematics & Statistics

Duquesne University

[august.delbert@gmail.com](mailto:august.delbert@gmail.com)

304-639-3498

Summer 2022 Data Science Intern Challenge Submission

**Question 1:** Given some sample data, write a program to answer the following: [click here to access the required data set](https://docs.google.com/spreadsheets/d/16i38oonuX1y1g7C_UAmiK9GkY7cS-64DfiDMNiR41LM/edit#gid=0)

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.

**Response:** I wrote a program using python to solve the question. The program is included with this submission and titled “[SneakerAnalysis.py](https://github.com/adelbert1/Shopify-Sneaker-Data-Set-Analysis/blob/main/SneakerAnalysis.py)”. The GitHub repository which contains all materials related to this question is available at: <https://github.com/adelbert1/Shopify-Sneaker-Data-Set-Analysis.git>. More details are provided below.

**Q1 Part A:** Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.

**Response:** My first step before getting deeper into any analysis is to familiarize myself with the data. After a review of the distributions of each variable and the basic descriptive statistics (mean, median, min, max etc.), two things stand out: 1) There are some unusual patterns in the data that don’t align with characteristics of the attributes I would expect based on the information provided in the prompt; 2) The most basic unit of interest (the number of pairs of sneakers per order) is not considered in the initial analysis. The fact that a single order can contain multiple pairs of sneakers appears to have been overlooked.

Note: My initial review of the visualizations of the basic descriptive statistics is saved in the repository as “[SneakerAnalysis\_Distribution\_withOutliers](https://github.com/adelbert1/Shopify-Sneaker-Data-Set-Analysis/blob/main/SneakerAnalysis_Distributions_withOutliers.pdf)” and “[SneakerAnalysis\_Distribution\_excludeOutliers](https://github.com/adelbert1/Shopify-Sneaker-Data-Set-Analysis/blob/main/SneakerAnalysis_Distributions_excludeOutliers.pdf)”. Outliers were excluded using the Median and Interquantile Deviation Method (IQD).

1. Regarding unusual patterns, characteristics, and attributes:

* For shop\_id = 78, the retail value of a pair of sneakers sold at their shop is $25,725. There are 46 purchases from this shop over a period of 30 days. Of these 46 purchases, 18 were made with cash. Before getting into any more sophisticated analysis, this stands out as not aligning with conventional reality/wisdom/expectations regarding sneaker sales. I would flag this as an area for closer review with the following considerations:
  + Perhaps this is a typo? The price includes two “25”s, so it’s possible someone may have accidentally double-typed this if they were in a rush with data entry.
  + Perhaps these are designer, high-demand shoes? Perhaps it is money laundering? Either way, the outlier is affecting our analysis and I would seek some insight on to make sure there was a reasonable explanation as to why the value would be so high.

1. Regarding the method of analysis:

* Based on the prompt, the intention is to analyze the average order value (AOV) across the 100 sneaker shops with the expectation that it should fall into alignment with conventional wisdom about the average cost/value of a pair of sneakers.
* The initial calculation (AOV = $3145.13) is simply an average of all order values.
* Some considerations for the analysis:
  + For each piece of data, shop\_id is not unique. Some shops are placing more orders than others. Since the AOV was calculated by averaging all order values for all shops, including multiple orders from a single shop, the shops with the most orders will skew the average value of a pair of sneakers.
  + A single order can include multiple pairs of sneakers. Without attention to the number of sneakers in a single order, multiple pairs are being taken as a single, representative unit. Without having that context, the use of AOV can be perceived as an inflated estimate of the cost of a pair of sneakers.
  + There are some extreme outliers in the data (determined visually and with IQD). These outliers are inflating the AOV relative to the median.
* From here, we may need to ask some clarifying questions. The prompt says that the analyst wants to focus on order values, but then goes on to consider order value relative to sneaker value, which is a related, but different concept altogether. What is the purpose/goal of analyzing the AOV? Is it to understand AOV for the sake of AOV, or could it instead be to get a general estimate of the value of a pair of sneakers? Some insight into this would help to select the most appropriate metric(s).
* Since the goal of the analysis is to focus on the AOV, I would recommend looking at the inter-quartile range (IQR), meaning the range of values from the 25th to the 75th percentile. This would offer a way to get around the troublesome outliers and provide a better sense of the distribution of values around the median.
* To ensure a more accurate analysis, I would recommend that the actual unit of interest, Average Sneaker Value (ASV), be the metric of focus. Average sneaker value can be determined by summing the total value of all sneaker orders and dividing that by the

**Q1 Part B:** What metric would you report for this dataset?

**Response:** If I could only pick one metric, it would be the inter-quartile range including the 25th, 50th, and 75th percentile values. This would also lend itself well to a boxplot, which could help someone quickly visualize the median and its spread to the 25th and 75th percentiles.

If it was appropriate, I would offer a few more metrics for both Average Order Value and Average Sneaker Value per order (min, max, median, percentiles, etc.) in a simple visualization like a boxplot to provide some additional helpful context.

For example:

**Q1 Part C:** What is its value?

|  |  |
| --- | --- |
| Order Amounts | Inter-Quantile Range (IQR) |
| 25th % | $163 |
| Median (50th %) | $284 |
| 75th % | $390 |

Chart, box and whisker chart

Description automatically generated

Note: The boxplots were generated using the python file “[SneakerAnalysis.py](https://github.com/adelbert1/Shopify-Sneaker-Data-Set-Analysis/blob/main/SneakerAnalysis.py)”. In most cases, I would use something like JMP to develop the boxplots since it can be quicker, but I opted to write this script since the prompt asked for a program.

**Question 2:** For this question you’ll need to use SQL. [Follow this link](https://www.w3schools.com/SQL/TRYSQL.ASP?FILENAME=TRYSQL_SELECT_ALL) to access the data set required for the challenge. Please use queries to answer the following questions. Paste your queries along with your final numerical answers below.

**Q2.A:** How many orders were shipped by Speedy Express in total?

**Response:** Query:

SELECT S.ShipperName, COUNT(O.OrderID) as TotalOrders

FROM Shippers S, Orders O

WHERE S.ShipperID = O.ShipperID

AND S.ShipperName = 'Speedy Express'

Numerical Answer:

Speedy Express: 54

**Q2.B:** What is the last name of the employee with the most orders?

**Response:** Query:

SELECT LastName, MAX(TotalOrders) as HighestTotalOrders

FROM (

SELECT DISTINCT E.LastName, COUNT(O.OrderID) as TotalOrders

FROM Employees E, Orders O

WHERE E.EmployeeID = O.EmployeeID

GROUP BY E.LastName

ORDER BY COUNT(O.OrderID) DESC

)

Numerical Answer:

Peacock: 40

**Q2.C:** What product was ordered the most by customers in Germany?

**Response:** Query:

SELECT ProductName, MAX(TotalOrders) as HighestTotalOrders

FROM (

SELECT P.ProductName, COUNT(OD.OrderDetailID) as TotalOrders

FROM Customers C, Orders O, OrderDetails OD, Products P

WHERE P.ProductID = OD.ProductID AND OD.OrderID = O.OrderID AND O.CustomerID = C.CustomerID

AND C.Country = 'Germany'

GROUP BY P.ProductName

ORDER BY COUNT(OD.OrderDetailID) DESC

)

Numerical Answer:

Gorgonzola Telino: 5

Some Additional Thoughts on Q2.C:

When you look at the number of orders for each product, you see that there is a large number of unique products (N = #), but that the median number of orders across all products is just 1. Noting that the maximum number of orders for any given product is 5, the volume of orders struck me as being relatively low. (I was asking myself, is five orders of a small food item that much more significant than 1 or 2 orders of another small food item across the entirety of a country? I admittedly have a lack of knowledge around this particular market sector; however, it is something I would seek to understand.)

Alternatively, when I look at the total quantity of each product that has been sold (recognizing that a single order for a product may include a quantity of that product greater than 1, I see that the total quantities sold increases substantially for each product and that there is a higher degree of variance amongst the quantities of products sold across all orders. For example,

With this in mind, I would consider providing information about both the product with the highest number of orders and the product with the highest total quantity sold (assuming this was within the bounds Shopify policy when working with a client/stakeholder), or I would seek clarification on the intention of this query to ensure I’m providing the client with the information that is truly most useful to them.