Applicant Name: **Geoffrey Gordon Ashbrook**
Position: Data Science Intern
Company: Shopify
Date: 2020.09.03
Source: Lambda School
Application Page: Here
Online Version of This: Here
Online Version of All Materials: Here

Shoe store AOV Analysis

Question 1: Given some sample data, write a program to answer the following: click here to access the required data set

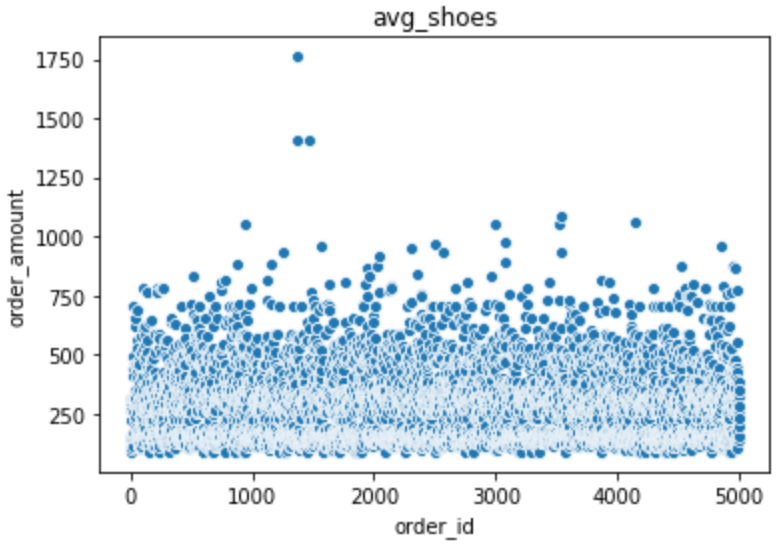
> On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of $3145.13. Given that we know these shops are selling sneakers, a relatively affordable item, something seems wrong with our analysis.
>
> a. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.
> b. What metric would you report for this dataset?
> c. What is its value?

Answer:

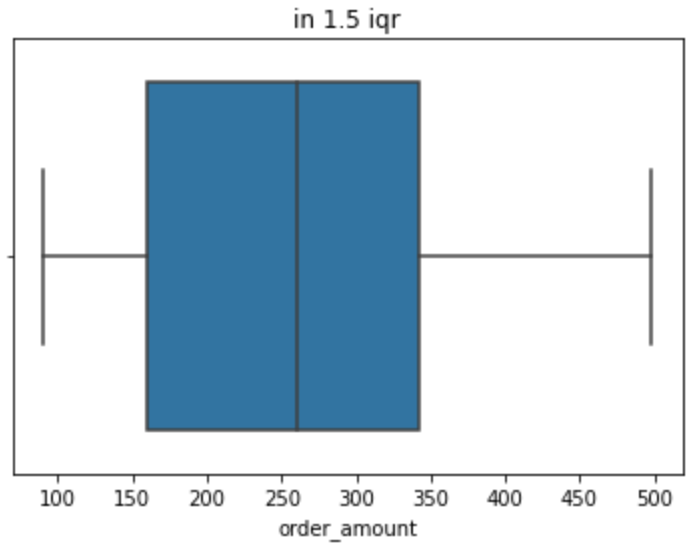
The Full Analysis Colab Notebook is here: https://colab.research.google.com/drive/1wjsghDJMOeADTvipP-57Ey3BSmEj2Ps3?usp=sharing

Answer Summary Below (From the first section of the full notebook):

Scatter Plot of All "Affordable Shoes" Orders



Box & Whisker Plot of 1.5 IQR of Order Amount (Box's Middle Line is Median)



Average Order Value: A Choice of Metrics

What metric and value would you report for this shoe store dataset?

Example Metrics:

- **mean** of All Data = $3145.128
- **median** of All Data = $284.0
- **mode** of All Data = $153.0

Full Question:

Look at orders data over a 30 day window. Out of 100 sneaker shops, where each shop sells one model of shoe, we want analysis of the average order value (AOV).

Initially we calculated an AOV of $3145.13. Given shops are selling **affordable** sneakers, something seems wrong with our analysis. What could be going wrong with our calculation?

Find a better way to evaluate this data. What **metric** would you report for this dataset? What is the value of this metric?

Provide your thought process & work.

(See Original Assignment Here)
https://docs.google.com/document/d/13VCtoyto9X1PZ74nPI4ZEDdb8hF8LAlcmLH1ZTHxKxE/edit?usp=sharing

---

Answer (Summary):

AOV = Median of 1.5 IQR (Interquartile Range) of order_amount = $260.00

Thought Process & Work (Summary): $260 is 7.2%

After using the standard statistical method 1.5_IQR (repeatable, and replicable, explained below) to remove 'outliers,' the resulting set of sales data looks much more promising than a raw mean for the following reasons. Within 1.5_IQR the mean and median differ only by .96, and the data still cover 99.3% of customers, 98% of shops, and 87.3% of orders. This is important because the more normally distributed and less skewed the data are, then the more compatible that data is with meaningful statistical analysis, and a normal distribution should include most of the data (also explained below).

For example, taking the median of the raw data (284) would be close to $260, but with the median and mean so far apart the data are very skewed. The overall median is likely **much better** than the overall mean but still of unknown reliability, and with no clarity of perception into what portion of the data that number describes (What good is a number, if you don't know what it refers to or how accurate it is, or how sound the data are, only knowing it comes from a very skewed distribution?). Additionally, on the chance that margins are thin for this segment of businesses, 284 USD and 260 USD may not be considered close at all.

The 1.5_IQR median on the other hand is much more clear. With a nearly identical mean and median in the 1.5_IQR data (still covering over 98% of shops and customers), we can be much more confident that the 1.5_IQR AOV number robustly describes an 'average' order, and we know much more about what specifically that number refers to within the data-set.

Sample Report Summary:

- mean using 1.5 IQR from all data = 260.9594
- median using 1.5 IQR from all data = 260.0
- mode using 1.5 IQR from all data = 153.0
- % of unique_order_id = 87.2625474905019
- % of unique_shop_id = 98.01980198019803
- % of unique customer id = 99.33774834437085

BUT

The isolation of this more normally distributed average customer, while looking indeed like the average shopper/order, represents a portion of customers that generated only 7.2% of sale dollars and 17% of items sold. This portion of orders may be your 'average customer & order placed' but it is not your average dollar earned or your average item-sold.

- % of sales = 7.2404002644342045
- % of_all_items_sold = 17.56418426802622

See below for a more complete analysis of how and why 1.5_IQR was chosen and an analysis of three likely categories of customers and shops, including one shop with one customer who make up %76.1 of sales revenue and %77.4 of sales volume.

Recommended actions based on this analysis:

- Keep your one big customer by any means (hire specialists to focus on and attend to them), and get more similar big customers if possible. Losing that one customer could sink your whole business.
- Find out what the other 'mystery' high-price customers are buying. Maybe have separate advertising and outreach for them.
- Try to increase the number of 'average' customers. They are probably the most stable and reliable part of your business, but they represent such a small % of your income (so small that they may be within a rounding error of the other customer groups in a given year). Some business strategies would recommend ditching these customers to focus on high-paying customers, but I think that is a high-risk strategy and ethically dubious in a broader perspective of your business' 'stakeholders.'

The aim here has been to provide analysis that is 1. statistically meaningful, 2. repeatable (given the same data and equations/notebook), 3. reproducible (across different analysts, not based on any arbitrary or intuitive decisions), 4. understandable, and 5. actionable.