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 Document: Here

## Contents:

## Two Essay Questions:

*1. Why are you interested in doing an internship with Shopify?*

This Shopify internship is a chance for me to realize my goal of training at a serious and well organized company from which I can learn competent practices, both those particular to my role, and those involving the culture and norms within a data driven business.

Helping society benefits everyone. Shopify's services appeal to me because I have long advocated for using data to help society. In the first chapter of my work-life I worked as a science teacher at companies and schools because of my belief that education and communication are of high value to society. Now I feel I can help society with Data Science by helping Small to Medium sized Enterprises (SMEs) to use and learn from their data. As a lover of the interdisciplinary and applied, helping employers make data driven decisions is a great place to apply my talents.

Using data to help society is important at all times, but in this age of covid-19 I am particularly motivated to find work that allows me to help the many vital SMEs that are the lifeblood of the international community. As the question was asked during and after both world wars "What did you do during the war?" It is important to me that during these times I do everything possible to stand for social justice and to empower as many people in their livelihoods as possible.

I am captivated by the type of work that the Data Science team does at Shopify. The DS challenge for this application was incredibly interesting and fun (diving into a complete story about a business from clues buried in data) and I am eager to continue working on this kind of fascinating analysis.

*2. Why are you the right person for this role? (This is a great place to tell us anything interesting about yourself that might not be indicated on your resume)*

Many know the classic scenario of the University classroom full of unfocused young students and a few 'older' students (the 30 and 40 year olds) who reliably attend, focus, take notes, talk with the professor, engage with the material, ask serious questions, and perform well. I am that older student. I have work experience and team project experience. I am serious, focused, ready to engage, eager to learn, and ready to work.

A broad background is an asset. I have a broad background across the sciences (having been a science teacher) as well as computer science (having worked in AI research as an undergrad). I have also worked for a dozen small businesses over the years, and have tried to start my own more than once, which should help me to see from your clients' perspectives. As Shopify grows into new areas, from fulfillment and banking to multiple language integration, having an intern or employee who is actively interested in all these areas will be an asset.

The right mindset can make a world of difference. From my background in the boyscouts to team project work in Education, my experience has strongly impressed on me the importance of Agile type best practice and positive values.

Lastly, I am strongly motivated by helping others, which has been a common thread throughout my past career and life, and likewise I will be motivated to help your clients.  From my choice to become a science teacher, to my choice after the tsunami in Japan to change jobs and move into the affected areas, I am fully engaged when it comes to helping others. I truly believe in the work Shopify and their data science team are doing and am genuinely motivated to help your clients. I believe you need someone who will engage and follow through with the work. That is me.Two Essay Questions

## Two Technical Challenges

*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

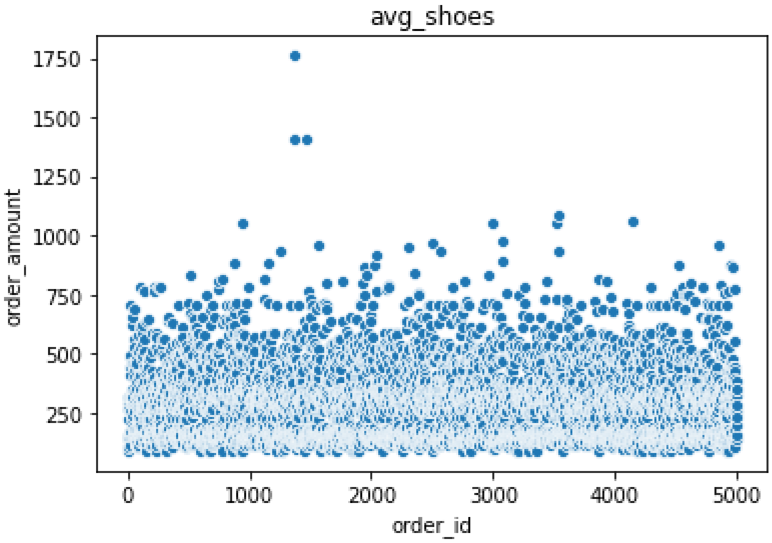
> On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of $3145.13. Given that we know these shops are selling sneakers,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

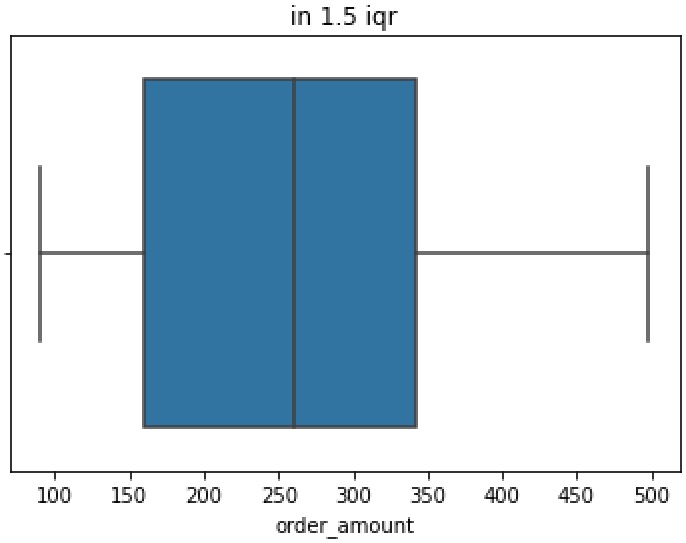
The Full Analysis Colab Notebook is here: https://colab.research.google.com/drive/1wjsghDJMOeADTvipP-57Ey3BSmEj2Ps3?usp=sharing

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.

*SQL*

**Question 2:** For this question you'll need to use SQL. Follow this link 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.

*a. How many orders were shipped by Speedy Express in total?*

(Query A)

```
Select count(*)
FROM
(
Select o.OrderID, o.ShipperID, s.ShipperName
FROM Shippers as s
INNER JOIN Orders as o
on o.shipperID = s.shipperID
WHERE ShipperName = "Speedy Express"
)
;
```

(Answer A)

54

*b. What is the last name of the employee with the most orders?*

(Query B)

```
SELECT LastName
FROM
(
SELECT LastName, MAX(employee_value_counts)
FROM
(
SELECT e.LastName, o.EmployeeID, COUNT(*) AS employee_value_counts
FROM Employees as e
LEFT JOIN Orders as o
ON e.EmployeeID = o.employeeID
GROUP BY e.EmployeeID
)
)
;
```

(Answer B)

Peacock

*c. What product was ordered the most by customers in Germany?*

(Query C)

```
SELECT ProductName
FROM
(
SELECT ProductName, MAX(product_value_counts)
FROM
(
SELECT c.Country, o.CustomerID, o.OrderID, p.ProductName,
od.ProductID, COUNT(ProductName) AS product_value_counts
FROM Customers as c
LEFT JOIN Orders as o
ON o.CustomerID = c.CustomerID
LEFT JOIN OrderDetails as od
ON od.OrderID = o.OrderID
LEFT JOIN Products as p
ON od.ProductID = p.ProductID
WHERE c.Country = "Germany"
GROUP BY od.ProductID
)
)
;
```

(Answer C)

Gorgonzola Telino

# Cover Letter

To the Hiring Team at Shopify,

I would love to contribute culturally and technically as a Data Scientist Intern  to your mission of making commerce better for everyone, lending an interdisciplinary perspective (1), and my work on AI ethics, to your teams and projects.

I began AI research in 2002, then, after going into Science Education for several years, I have been studying Data Science in Python and R since 2017. I recently completed a 9-month Data Science program (2) with a broad foundation (3) in statistics, linear algebra, data analysis, and machine learning and deployment from standard regression and classification models to Neural Networks. The combination of these skills with my background in science education, AI, and the sciences should make me a great asset to your work.

From my decade of experience in Education, I saw first hand how important project management, reviews, metrics, problem-identification, analysis, planning, goal cultivation, communication, coordination, and best practice are.

Work experience does not leave the same impression on everyone; to me work has left the impression of the importance of positive values, project best practice, and an 'Agile Mindset.' I was perhaps set up for this from my experience as a boy scout through to becoming an Eagle Scout which strongly stressed values and conduct. Making a daily habit of scrutinizing the boy-scout behavioral code, which on one level simply outlines decent behavior, I have come to see a significant overlap between "boy scout values" and 'The Agile Mindset,' specifically in the context of project-communication. There is a level of caring, conscientiousness, service, duty, and cooperation, which goes beyond the perfunctory but very necessary 'behaviors' of a SCRUM project. Being able to work with a team on projects is crucial and it is one of my key strengths.

My teaching experience also taught me how much I love working with people of all ages and backgrounds, and how important effective communication is, and that everyone can understand topics when the explanation is clear and the mood is welcoming. Here are two articles (4, 5) showing my interest in explaining technical topics to non-technical audiences. Working with diverse cross-functional-teams is my favorite part of projects.

As Shopify branches out into more areas including blockchain transactions, wholesale markets, in house fulfillment logistics, full remote work coordination, banking and finance management, multiple language integration (having linguistics, NLP & international work experience myself), and the ubiquitous cyber security landscape, having an intern or employee who is actively interested in all these areas will be an asset.

Thank you very much for your time and consideration. I look forward to having an opportunity to speak with members of your organization about your products and culture and about how I can best contribute.

Yours sincerely,
Geoffrey Gordon Ashbrook

(links separated for for txt-only version)

link 1: Data Science Book List: Recommended Background Reading
https://docs.google.com/document/d/1dDF40M5JjjrBsYYQbJplz3M738ktQBBYyNa6FXhzNFU/edit?usp=sharing

link 2: Lambda School's Very Brief Course Outline
https://lambdaschool.com/courses/data-science

link 3: Work Flow for Data Science & Machine Learning
https://medium.com/wooden-information/data-science-data-analysis-workflow-template-v1-3-11c9f6ff3e90

link 4: (Explaining Concepts Clearly to Any Reader) Time Leakage
https://medium.com/wooden-information/less-is-more-904427f568e0

link 5: (Explaining Concepts Clearly to Any Reader) NLP Embeddings
https://colab.research.google.com/drive/1n0QHVKLmjHhb1J0PVumoxq58-1OevP5b