# Fall 2022 Data Science Intern Challenge

Please complete the following questions, and provide your thought process/work. You can attach your work in a text file, link, etc. on the application page. Please ensure answers are easily visible for reviewers!

**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.

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

order\_amount = 3145.12 +/- 41282.53

total\_items = 8.78 +/- 116.32

There is a huge standard deviation relative to mean for both order\_amount and total\_items. From such large deviations, we can infer that there are outliers and large variance in the dataset. The standard error is quite high which signifies that the mean is likely an inaccurate representation of the true mean.

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

Median

1. **What is its value?**

I would report the median as the metric. We can observe that the median value including outliers is 284 and excluding outliers is 280, which shows the median is more robust than mean against the outliers.

**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.

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

select sum(od.Quantity) as total\_orders\_by\_speedy\_express

from OrderDetails od

join Orders o

on o.OrderID = od.OrderID

join Shippers s

on s.ShipperID = o.ShipperID

where ShipperName = 'Speedy Express';

Graphical user interface, application

Description automatically generated with medium confidence

*Result*: 3575

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

select e.LastName, sum(od.Quantity) as no\_of\_orders

from Employees e

join Orders o

on o.EmployeeID = e.EmployeeID

join OrderDetails od

on od.OrderID = o.OrderID

group by e.LastName

order by no\_of\_orders desc limit 1;

Graphical user interface, application

Description automatically generated

*Result*: Peacock

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

select p.ProductName, sum(od.Quantity) as total\_orders

from Products p

join OrderDetails od

on od.ProductID = p.ProductID

join Orders o

on o.OrderID = od.OrderID

join Customers c

on c.CustomerID = o.CustomerID

where c.Country = 'Germany'

group by p.ProductName

order by total\_orders desc limit 1;

Graphical user interface, text, application

Description automatically generated

*Result*: Boston Crab Meat