# Summer 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.

Answer:

For all orders of different products in multiple stores, if we simply sum the mean value, the data will be biased, because the data distribution is not necessarily normal distribution, and the left or right deviation will make the mean value can not well describe the distribution of the data median. Similar situations often occur when we seek the average salary for a region. Some managers with excessive salary will make the average salary falsely high, which can not well measure the salary of most people. Therefore, we should use the method of finding the median.

Chart

Description automatically generated

1. What metric would you report for this dataset?

Answer: Median

1. What is its value?

Answer: 284

**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?
2. What is the last name of the employee with the most orders?
3. What product was ordered the most by customers in Germany?

SQL Answers

select count(OrderID) as total\_speedy\_orders

from Orders o

left join Shippers s

on o.ShipperID = s.ShipperID

where ShipperName = ‘Speedy Express’

with t as (

select o.EmployeeID,

LastName,

count(OrderID) as total\_orders

from Orders o

left join Employees e

on o.EmployeeID = e.EmployeeID

group by o.EmployeeID, LastName

order by count(OrderID) desc)

select LastName

from t

where total\_orders in (select max(total\_orders) from t)

with t as (

select d.ProductID,

c.CustomerID,

sum(d.Quantity) as total\_orders

from OrderDetails d

join Orders o

on d.OrderID = o.OrderID

join Customers c

on o.CustomerID = c.CustomerID

where c.Country = 'Germany'

group by d.ProductID)

select t.ProductId,

p.ProductName

from t

join Products p

on t.ProductID = p.ProductID

where t.toal\_orders in (select max(total\_orders) from t)