

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](#)

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.**

Just by understanding the problem, it seems like there are outliers present in the “order value” data (these outliers may be orders with high amount) due to which the data is skewed to the right. Because of the presence of outliers, the calculated mean would be high, which is why, the AOV is \$3145.13 for sneaker orders.

A better way to evaluate this data would be by calculating the median of the order value. Median gives us the value present in the middle of the data, which often is used to replace the mean when there are outliers present in the data.

- b. What metric would you report for this dataset?**

I would use the median to report for this dataset instead of the mean.

- c. What is its value?**

Median of order value = 284

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:

```
SELECT
  count(o.OrderID) as total_orders
FROM
  Orders o
  join Shippers s on o.ShipperID = s.ShipperID
where
  s.ShipperName = 'Speedy Express';
```

Answer = 54

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

Query:

```
with max_count as (
  SELECT
    EmployeeID,
    count(OrderID) as cnt
  FROM
    Orders
  group by
    EmployeeID
)
SELECT
  e.LastName
from
  max_count m
  join Employees e on m.EmployeeID = e.EmployeeID
where
  m.cnt = (
    select
      max(cnt)
    from
      max_count
  );
```

Answer = Peacock

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

Query:

```
with max_order as (  
  select  
    od.ProductID,  
    count(o.OrderID) as cnt  
  from  
    orders o  
    join OrderDetails od on o.OrderID = od.OrderID  
    join Customers c on c.CustomerID = o.CustomerID  
  where  
    c.Country = 'Germany'  
  group by  
    ProductID  
)  
select  
  p.ProductName  
from  
  max_order m  
  join products p on m.ProductID = p.ProductID  
where  
  cnt = (  
    select  
      max(cnt)  
    from  
      max_order  
  );
```

Answer = Gorgonzola Telino