Winter 2021 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.

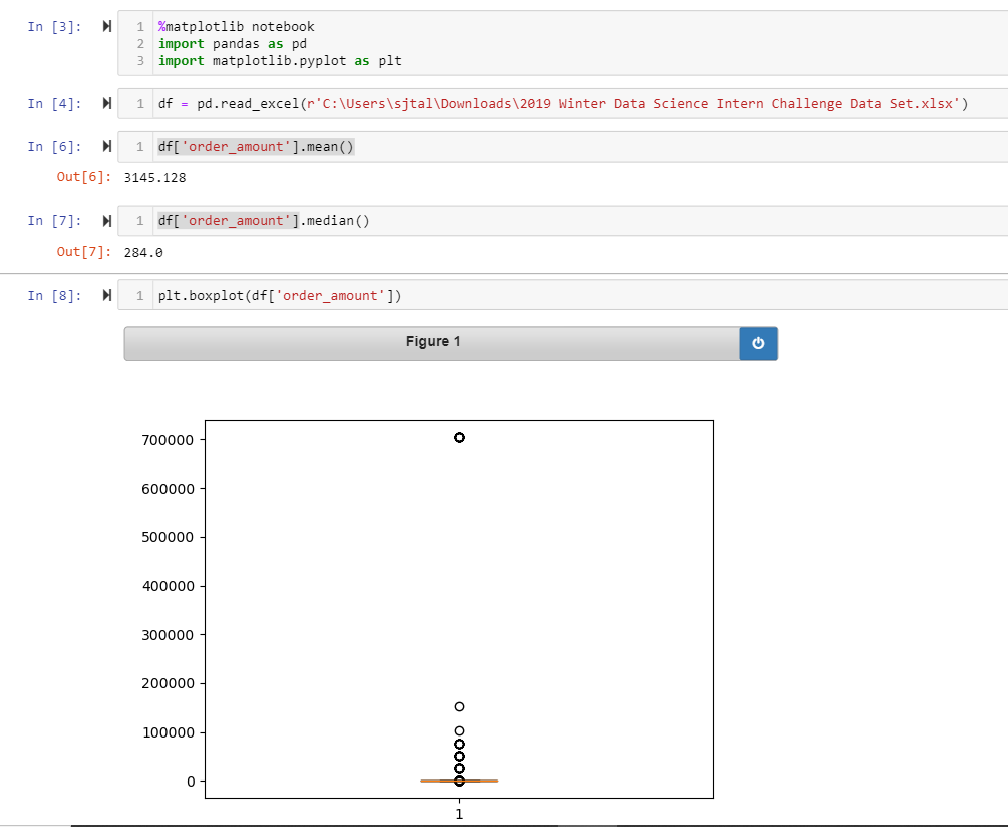
**ANSWER:** When a histogram and a boxplot is created from the order\_amount data, it is clearly seen that outlier data present in order amounts that differ widely away from the ‘average’ is affecting the reported amount, (In the spreadsheet when you FILTER the ORDER\_AMOUNT these numbers are visible too). In such a case the mean does not really describe the information we are seeking. The median is a better statistic that reflects the ‘average’.

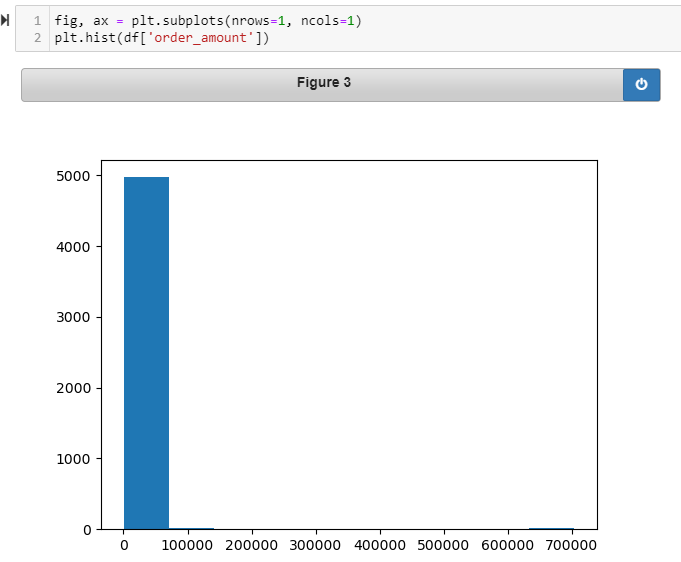
1. What metric would you report for this dataset?

**ANSWER: Median**

1. What is its value?

**ANSWER: 284**





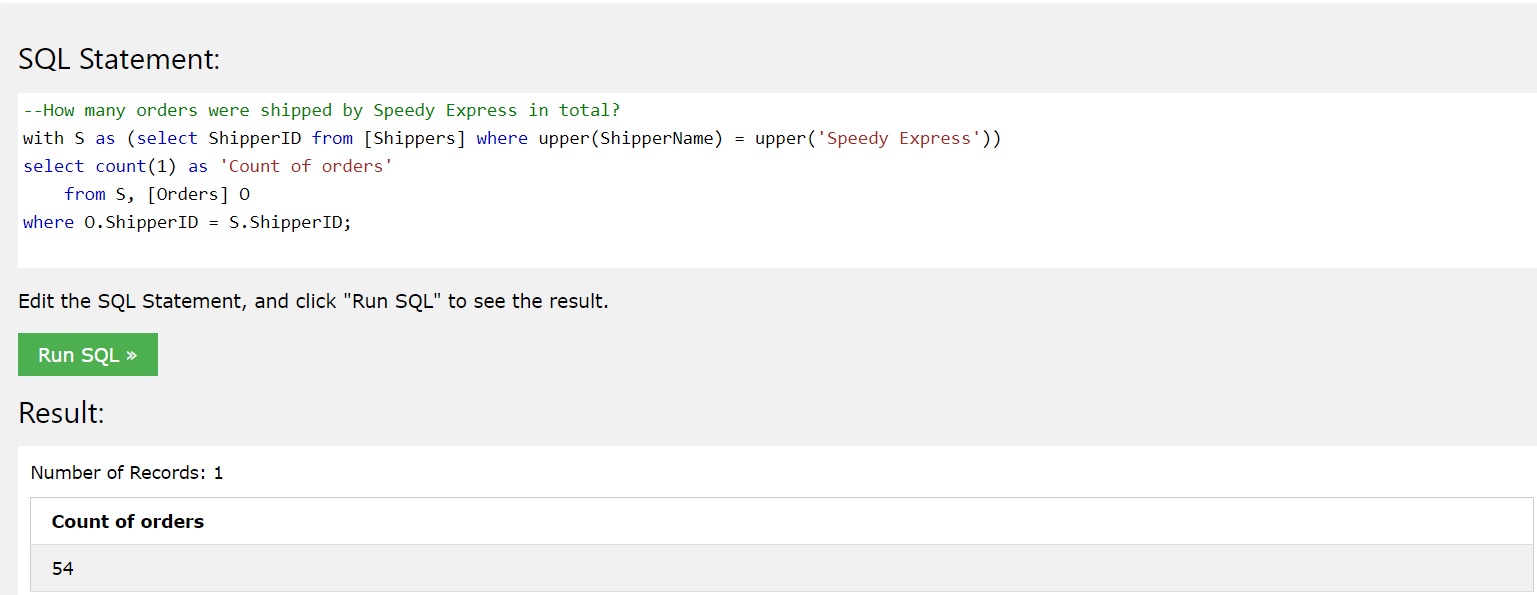
**Question 2:** For this question you’ll need to use SQL. [Follow this link](https://www.w3schools.com/SQL/TRYSQL.ASP?FILENAME=TRYSQL_SELECT_ALL) to access the data set required for the challenge. Please use queries to answer the following questions. Paste your queries along with your final numerical 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?

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

with S as (select ShipperID from [Shippers] where upper(ShipperName) = upper('Speedy Express'))  
select count(1) as 'Count of orders'  
 from S, [Orders] O  
where O.ShipperID = S.ShipperID;

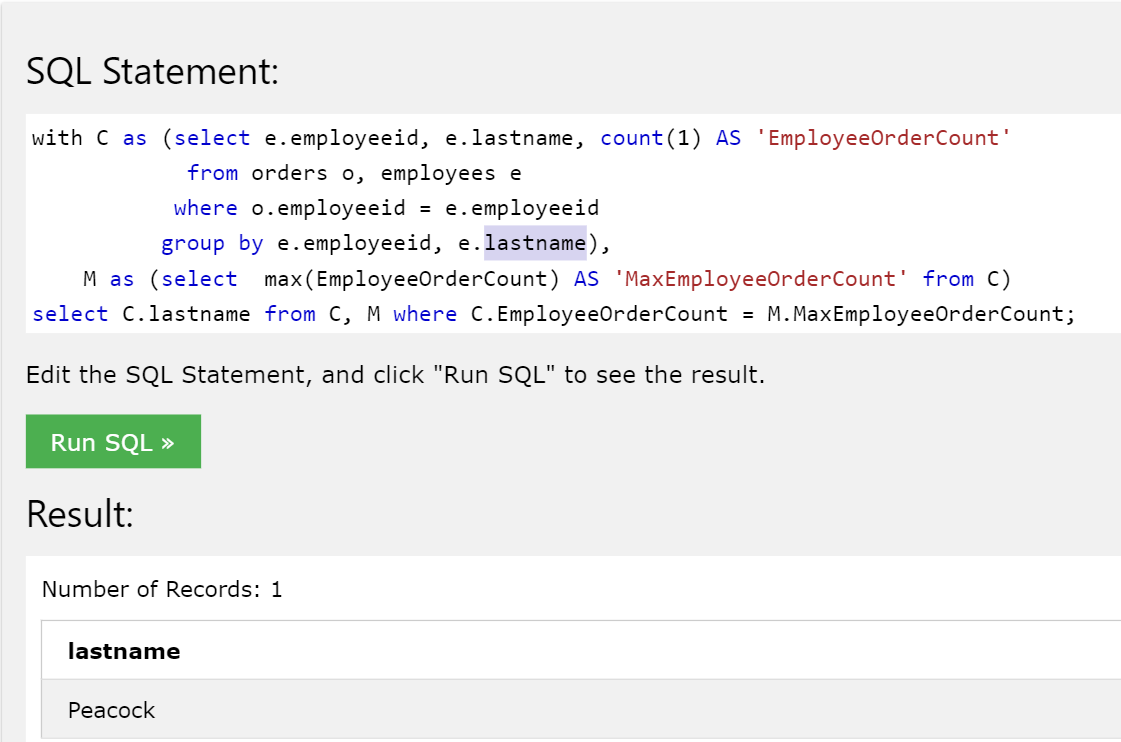
**Answer : 54**



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

with C as (select e.employeeid, e.lastname, count(1) AS 'EmployeeOrderCount'  
 from orders o, employees e  
 where o.employeeid = e.employeeid  
 group by e.employeeid, e.lastname),  
 M as (select max(EmployeeOrderCount) AS 'MaxEmployeeOrderCount' from C)  
select C.lastname from C, M where C.EmployeeOrderCount = M.MaxEmployeeOrderCount;

**Answer: Peacock**



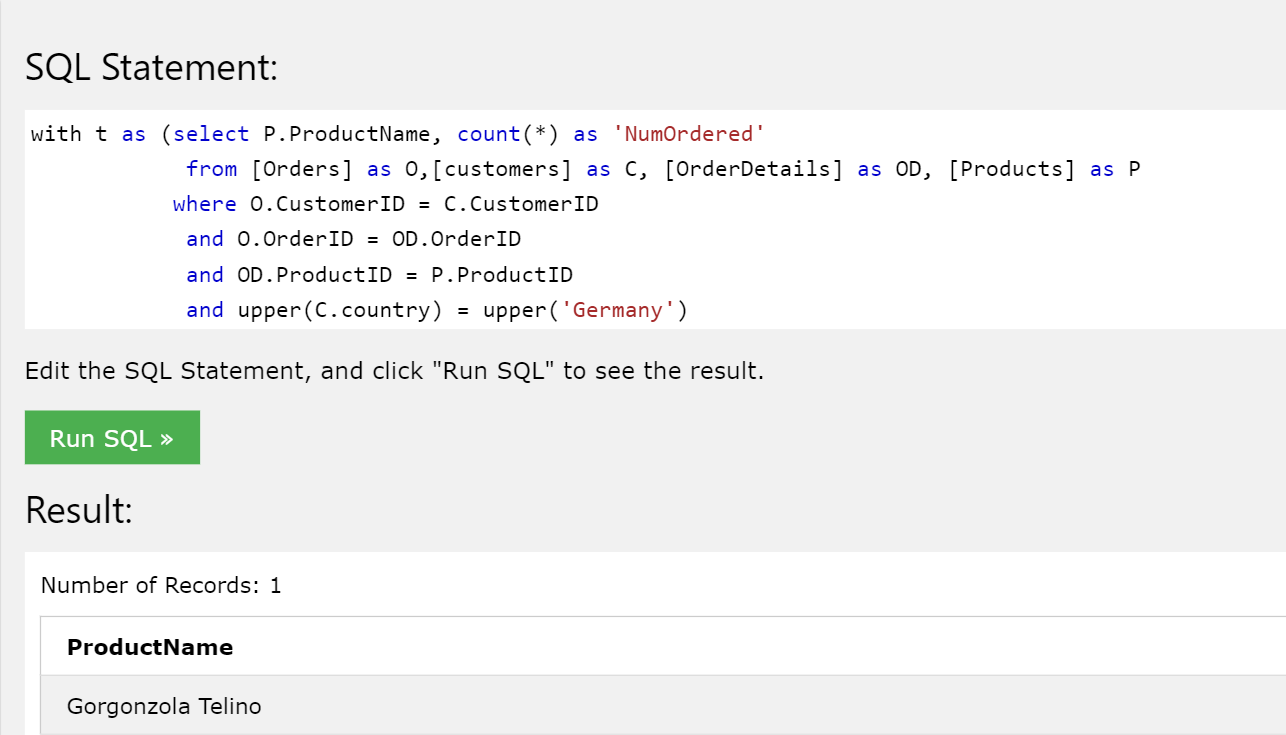
OPTION B: ANALYTIC FUNCTION WITH DISTINCT : DISTINCT FIRST\_VALUE(employeeid) OVER (PARTITION BY employeeid ORDER BY EmployeeOrderCount) will provide the same result.

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

with t as (select P.ProductName, count(\*) as 'NumOrdered'  
 from [Orders] as O,[customers] as C, [OrderDetails] as OD, [Products] as P  
 where O.CustomerID = C.CustomerID  
 and O.OrderID = OD.OrderID  
 and OD.ProductID = P.ProductID  
 and upper(C.country) = upper('Germany')  
 group by P.ProductName  
 order by NumOrdered desc)

select ProductName from t limit 1;

**ANSWER : GORGONZOLA TELINO**



**OPTION B:**

with t as (select P.ProductName, count(\*) as 'NumOrdered'

from [Orders] as O,[customers] as C, [OrderDetails] as OD, [Products] as P

where O.CustomerID = C.CustomerID

and O.OrderID = OD.OrderID

and OD.ProductID = P.ProductID

and upper(C.country) = upper('Germany')

group by P.ProductName

order by NumOrdered desc),

mp as (select max(NumOrdered) as 'MaxNumOrdered' from t)

select ProductName from t, mp

where t.NumOrdered = mp.MaxNumOrdered;