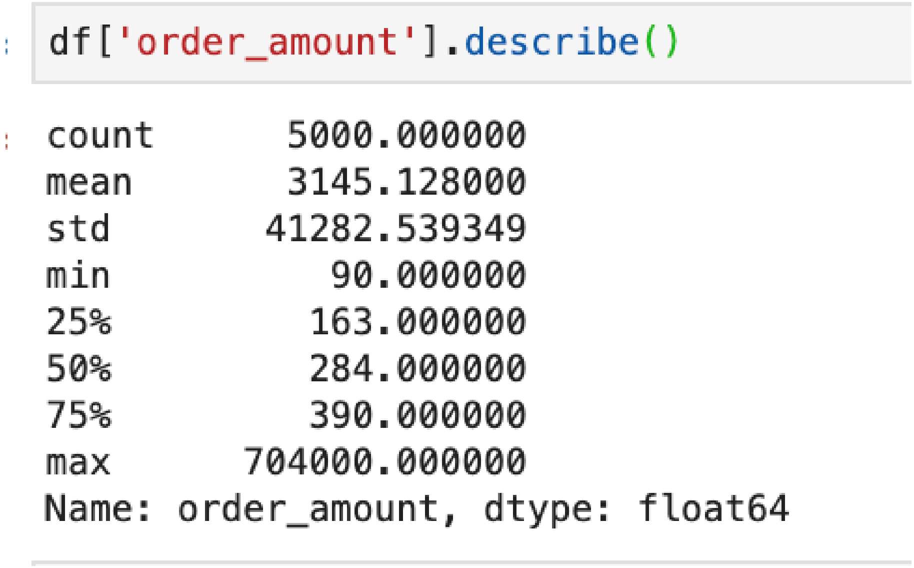
Winter 2021 Data Science Intern Challenge

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

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.

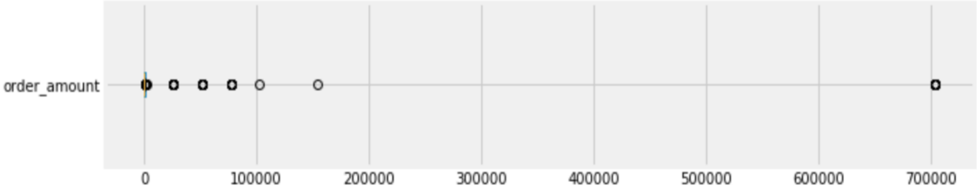
a. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.
Looking at the description of the data, seems to have outliers. Cleaning the data by removing the outliers will be a better way to evaluate the data.



```
df['order_amount'].describe()
```

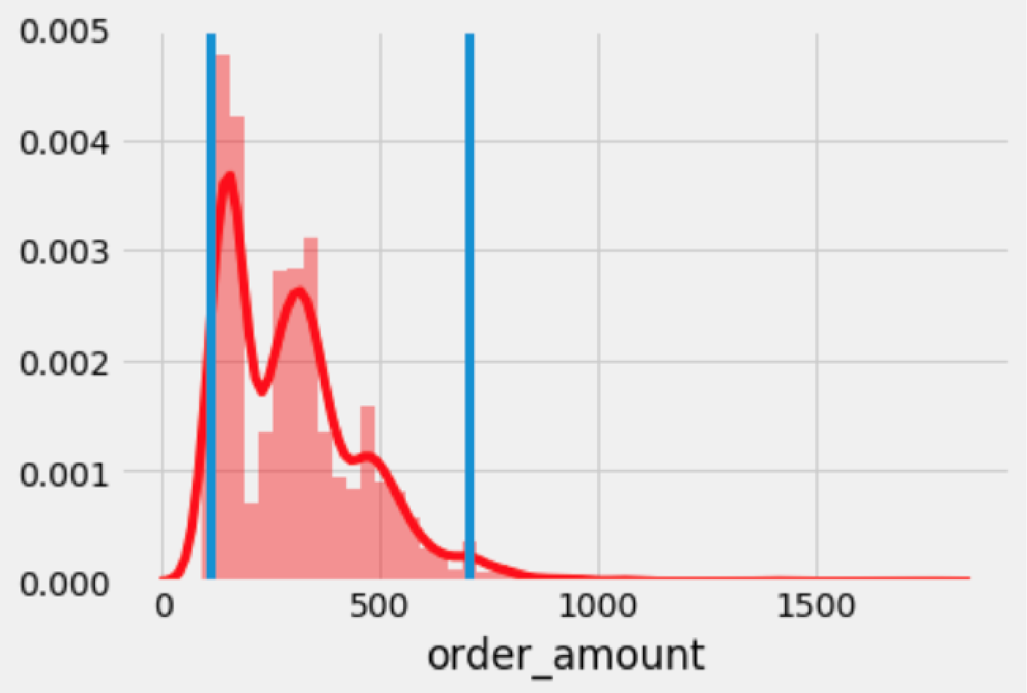
```
count       5000.000000
mean        3145.128000
std        41282.539349
min           90.000000
25%          163.000000
50%          284.000000
75%          390.000000
max       704000.000000
Name: order_amount, dtype: float64
```

Box plot distribution of the data:



b. What metric would you report for this dataset?

Remove the outliers in the order_amount. Created a distribution plot with 95 percentile window of the data.



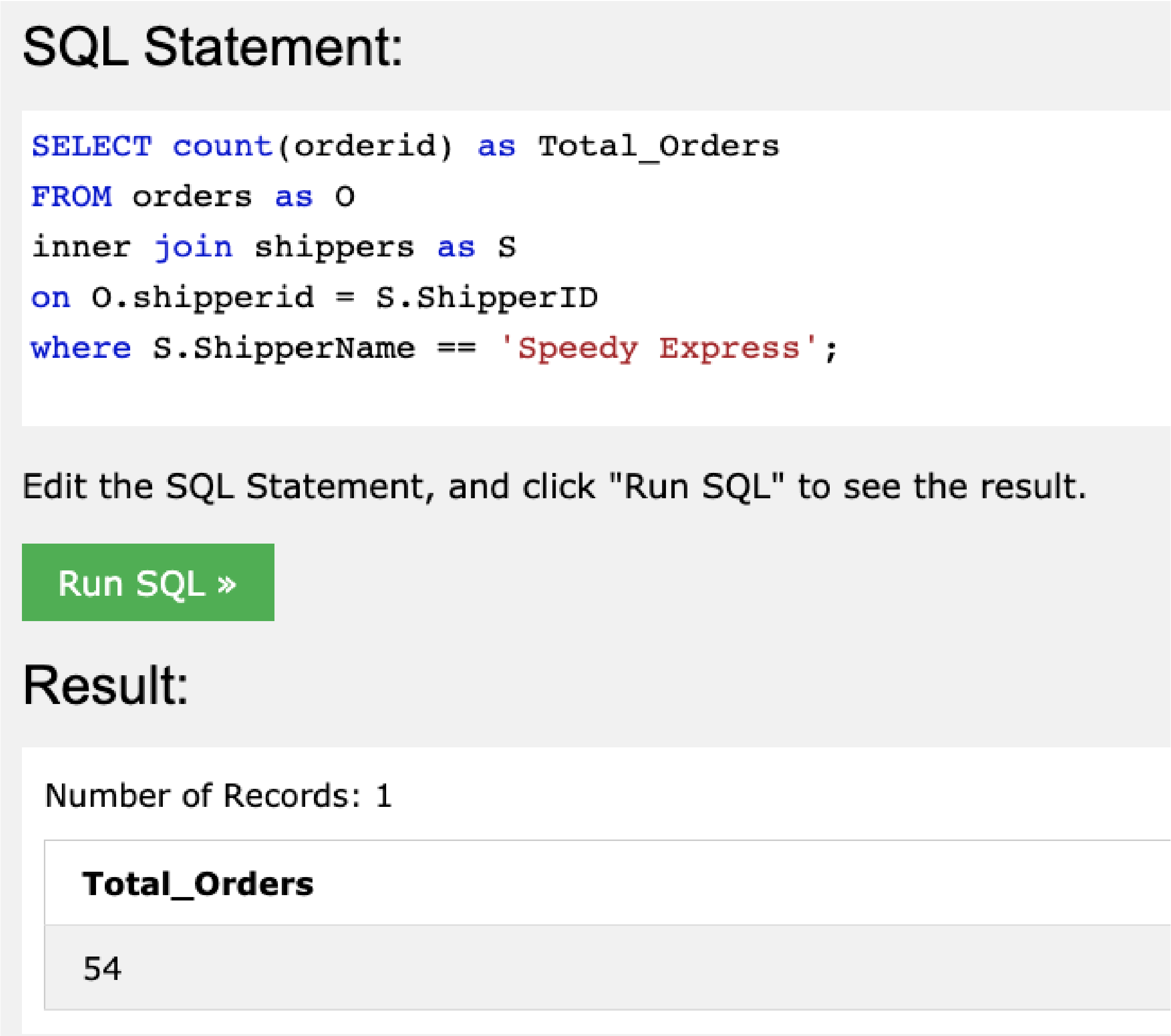
c. What is its value?

Considering the order amount from 80 to 750, the Average Order Value came to $294.91

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?
54

```
SELECT count(orderid) as Total_Orders
 FROM orders as O
inner join shippers as S
on O.shipperid = S.ShipperID
where S.ShipperName == 'Speedy Express';
```



SQL Statement:

```
SELECT count(orderid) as Total_Orders
FROM orders as O
inner join shippers as S
on O.shipperid = S.ShipperID
where S.ShipperName == 'Speedy Express';
```

Edit the SQL Statement, and click "Run SQL" to see the result.

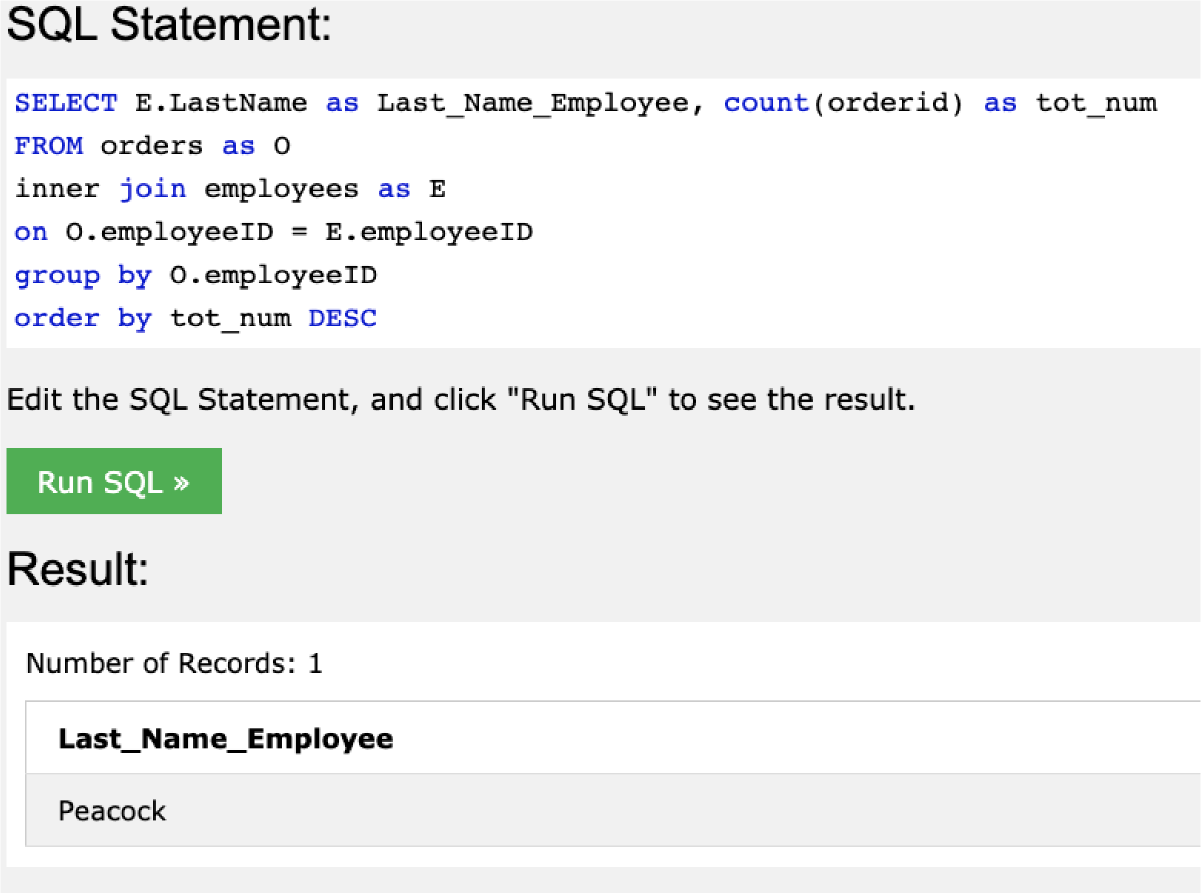
Result:

Number of Records: 1

Total_Orders
54

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

```
SELECT Last_Name_Employee from (
SELECT E.LastName as Last_Name_Employee, count(orderid) as tot_num
FROM orders as O
inner join employees as E
on O.employeeID = E.employeeID
group by O.employeeID
order by tot_num DESC
limit 1 );
```



SQL Statement:

```
SELECT E.LastName as Last_Name_Employee, count(orderid) as tot_num
FROM orders as O
inner join employees as E
on O.employeeID = E.employeeID
group by O.employeeID
order by tot_num DESC
```

Edit the SQL Statement, and click "Run SQL" to see the result.

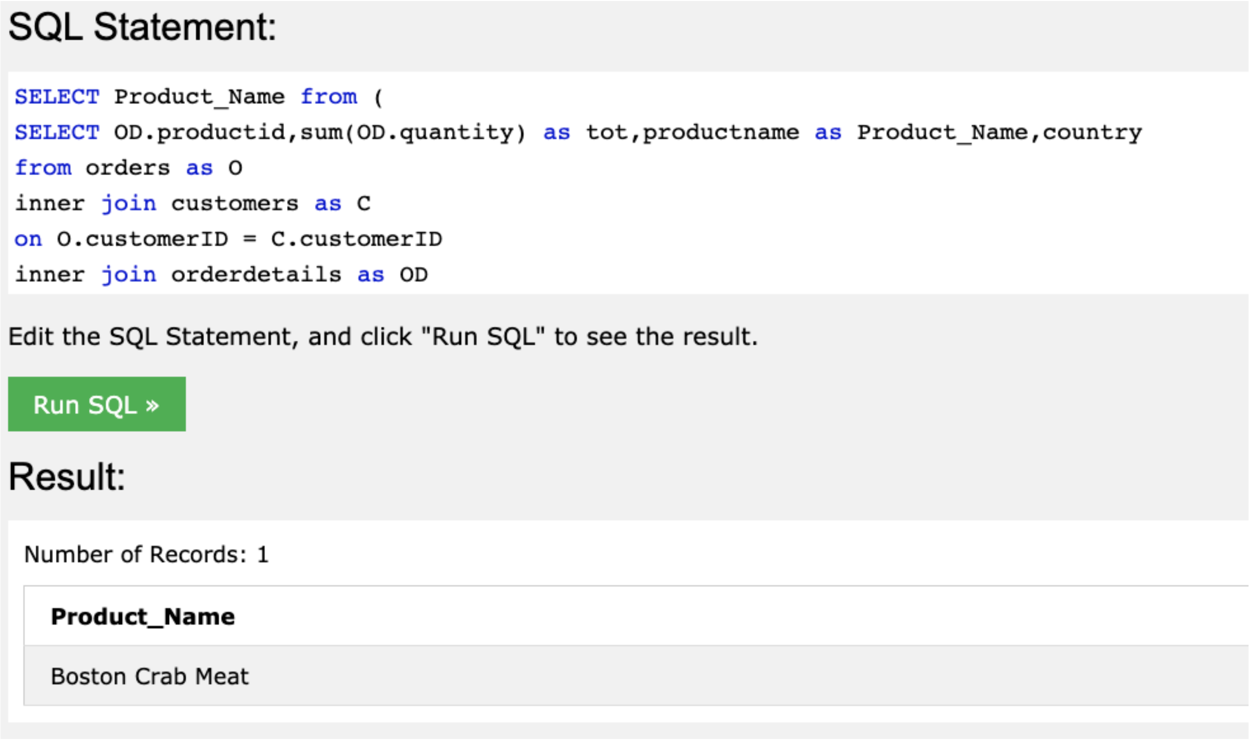
Result:

Number of Records: 1

Last_Name_Employee
Peacock

c. What product was ordered the most by customers in Germany?
Product_Name
Boston Crab Meat

```
SELECT Product_Name from (
SELECT OD.productid,sum(OD.quantity) as tot,productname as Product_Name,country
from orders as O
inner join customers as C
on O.customerID = C.customerID
inner join orderdetails as OD
on O.orderid = OD.orderid
inner join products as P
on OD.productID = P.productID
where C.country = 'Germany'
group by OD.productid
order by tot DESC
limit 1
);
```



SQL Statement:

```
SELECT Product_Name from (
SELECT OD.productid,sum(OD.quantity) as tot,productname as Product_Name,country
from orders as O
inner join customers as C
on O.customerID = C.customerID
inner join orderdetails as OD
```

Edit the SQL Statement, and click "Run SQL" to see the result.

Result:

Number of Records: 1

Product_Name
Boston Crab Meat