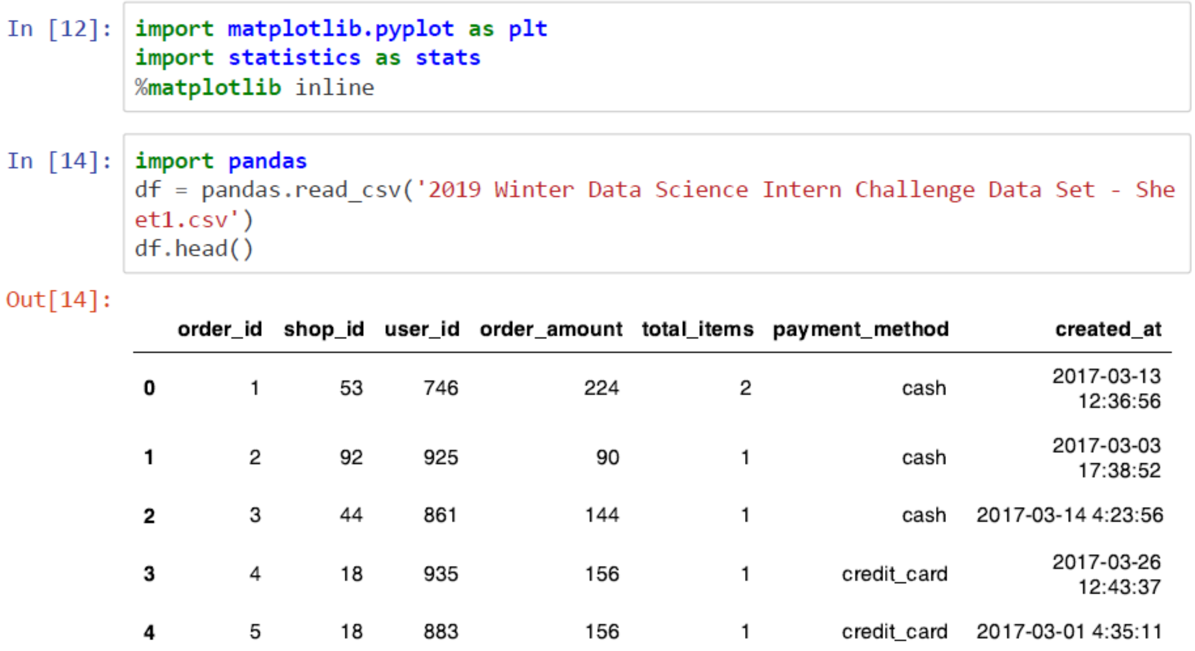
Fall 2021 Data Science Intern Challenge

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.
> b. What metric would you report for this dataset?
> c. What is its value?

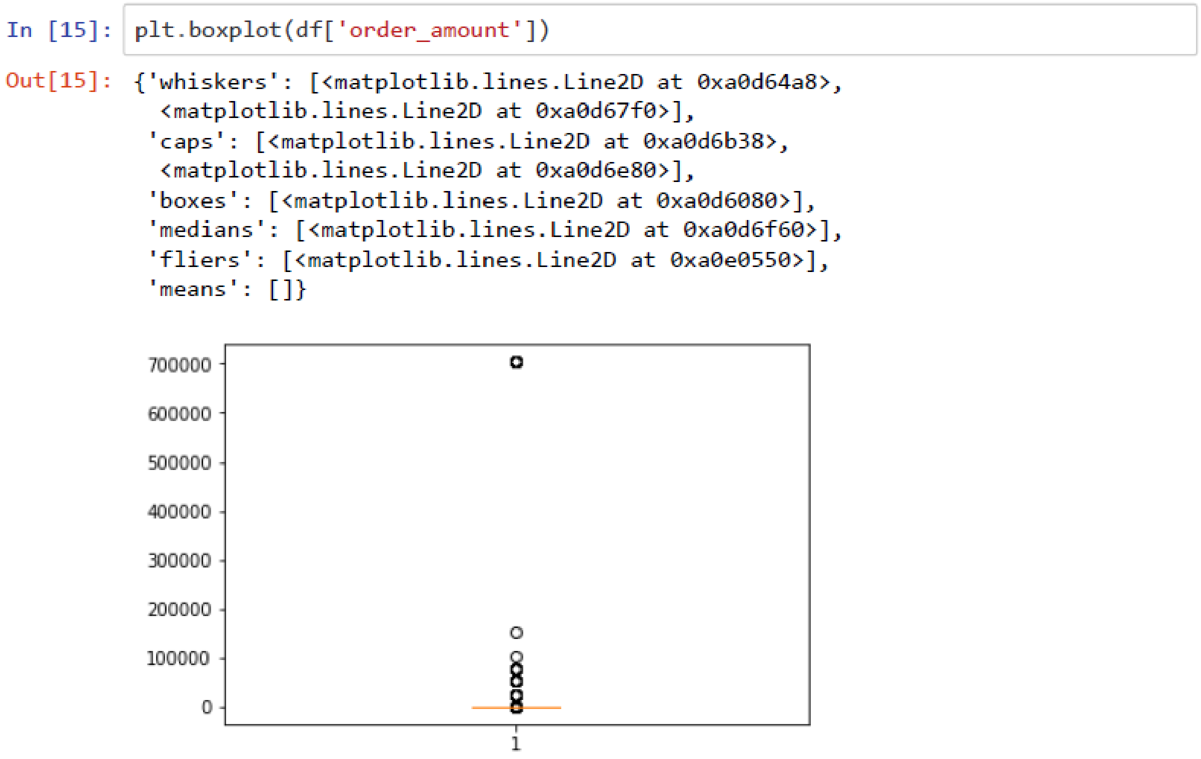


```
In [12]: import matplotlib.pyplot as plt
         import statistics as stats
         %matplotlib inline
```

```
In [14]: import pandas
         df = pandas.read_csv('2019 Winter Data Science Intern Challenge Data Set - She
         et1.csv')
         df.head()
```

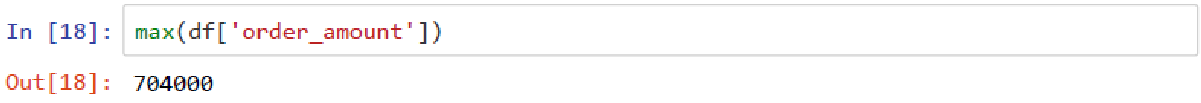
Out[14]:

	order_id	shop_id	user_id	order_amount	total_items	payment_method	created_at
0	1	53	746	224	2	cash	2017-03-13 12:36:56
1	2	92	925	90	1	cash	2017-03-03 17:38:52
2	3	44	861	144	1	cash	2017-03-14 4:23:56
3	4	18	935	156	1	credit_card	2017-03-26 12:43:37
4	5	18	883	156	1	credit_card	2017-03-01 4:35:11



```
In [15]: plt.boxplot(df['order_amount'])

Out[15]: {'whiskers': [<matplotlib.lines.Line2D at 0xa0d64a8>,
           <matplotlib.lines.Line2D at 0xa0d67f0>],
          'caps': [<matplotlib.lines.Line2D at 0xa0d6b38>,
           <matplotlib.lines.Line2D at 0xa0d6e80>],
          'boxes': [<matplotlib.lines.Line2D at 0xa0d6080>],
          'medians': [<matplotlib.lines.Line2D at 0xa0d6f60>],
          'fliers': [<matplotlib.lines.Line2D at 0xa0e0550>],
          'means': []}
```

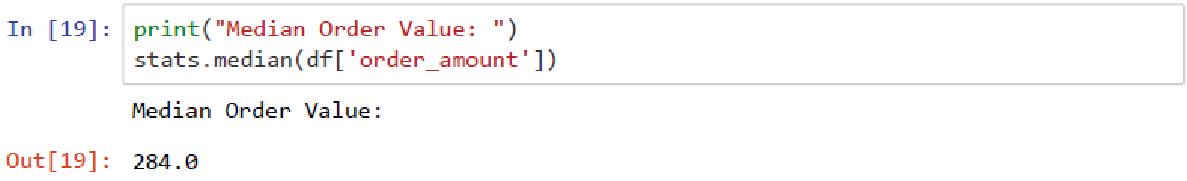


```
In [18]: max(df['order_amount'])

Out[18]: 704000
```

a. From the boxplot, we can observe that there are outliers in the data, and the maximum price is $704,000, which make the distribution skewed. We need to make sure if these values are an error in the data entry or not. But assuming that the data is taken from the transaction record in an ERP system, then we assume that these values are correct. In a skewed distribution like this, it is better to use a median value than an average value.

b. In a data with skewed distribution like this, it is better to use the Median Order Value than the Average Order Value. So instead of using the mean of the order_amount column, we will use its median.



```
In [19]: print("Median Order Value: ")
         stats.median(df['order_amount'])

         Median Order Value:

Out[19]: 284.0
```

c. The median order value is 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?
SQL code:

```
SELECT COUNT(OrderID) FROM orders WHERE ShipperID = 1;
```

Answer:
There were **54** orders in total.

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

```
SELECT LastName FROM Employees WHERE EmployeeID =
(SELECT EmployeeID FROM Orders
GROUP BY EmployeeID ORDER BY COUNT(*) DESC
LIMIT 1);
```

Answer:
Peacock

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

```
SELECT ProductName FROM Products WHERE ProductID = (SELECT ProductID
FROM (SELECT ProductID,SUM(Quantity) FROM OrderDetails WHERE OrderID IN
(SELECT OrderID FROM Orders WHERE CustomerID IN (SELECT CustomerID
FROM [Customers] WHERE Country = 'Germany')) GROUP BY ProductID ORDER
BY SUM(Quantity) DESC LIMIT 1))
```

Answer:
Boston Crab Meat