```
# import libraries to work with the data
import pandas as pd
import numpy as np
import math
import csv
import matplotlib.pyplot as plt
```

```
# importing the data and having a snapshot of it
path = '/content/2019 Winter Data Science Intern Challenge Data Set - Sheet1.csv'
data = pd.read_csv(path, encoding="utf-8")
data.head()
```

	order_id	shop_id	user_id	order_amount	total_items	payment_method	create
0	1	53	746	224	2	cash	2017-03-13 12:3
1	2	92	925	90	1	cash	2017-03-03 17:3
2	3	44	861	144	1	cash	2017-03-14 4:2
3	4	18	935	156	1	credit_card	2017-03-26 12:4
4	5	18	883	156	1	credit_card	2017-03-01 4:3

```
# to get the information of the data
data.describe()
```

	order_id	shop_id	user_id	order_amount	total_items
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.00000
mean	2500.500000	50.078800	849.092400	3145.128000	8.78720
std	1443.520003	29.006118	87.798982	41282.539349	116.32032
min	1.000000	1.000000	607.000000	90.000000	1.00000
25%	1250.750000	24.000000	775.000000	163.000000	1.00000
50%	2500.500000	50.000000	849.000000	284.000000	2.00000
75%	3750.250000	75.000000	925.000000	390.000000	3.00000
max	5000.000000	100.000000	999.000000	704000.000000	2000.00000

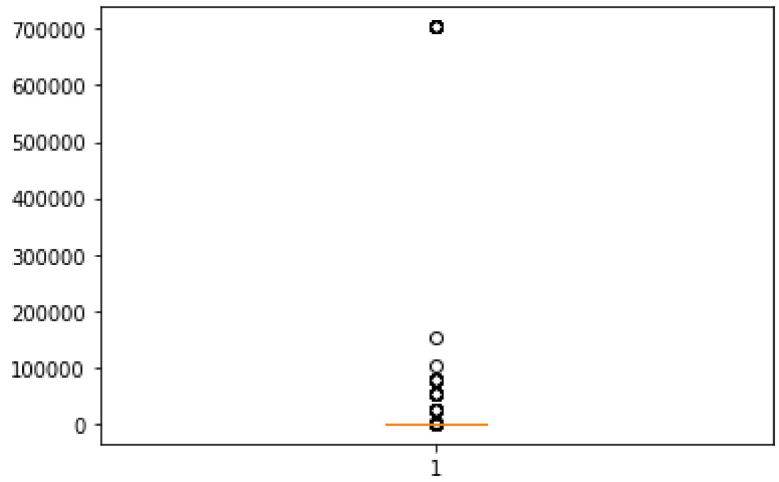
```
data.duplicated().sum()
```

```
0
```

1a) From the above description of the data set it could be found that there are no missing values and no duplicate values and that the mean of order_amount is given as 3145.13 and the standard deviation as 41282 which is a pretty high value. The mean is sensitive to outliers the wrong value might be due to the presence of outliers. Hence from the above statements we can say that mean is not a good metric to analyse and calculate AOV.

```
plt.boxplot(x = data['order_amount'])
```

```
{'boxes': [<matplotlib.lines.Line2D at 0x7f40ee982a50>],
 'caps': [<matplotlib.lines.Line2D at 0x7f40ee985ad0>,
  <matplotlib.lines.Line2D at 0x7f40ee98d050>],
 'fliers': [<matplotlib.lines.Line2D at 0x7f40ee98db10>],
 'means': [],
 'medians': [<matplotlib.lines.Line2D at 0x7f40ee98d5d0>],
 'whiskers': [<matplotlib.lines.Line2D at 0x7f40ee985050>,
  <matplotlib.lines.Line2D at 0x7f40ee985590>]}
```




```
# Calculating the mean and median of the order_amount
med = data['order_amount'].median()
print("The median value:", med)
mod1 = data['order_amount'].mode()
print("The mode value:", mod1)
```

```
The median value: 284.0
The mode value: 0    153
dtype: int64
```

1b) The metrics that we can use to calculate the correct value of AOV are mean and median. By using median we get a value of 284 dollars but we can see that our data is skewed, hence median would not be a very good metric to analyze the value of AOV. By using mode we get a value of 153$ which is in the generic price range of sneakers.

1c) Correct AOV : $153

✓ 0s completed at 8:39 PM

Shopify Data Science Intern Challenge

2a) How many orders were shipped by Speedy Express in total?

```
SELECT COUNT(DISTINCT OrderID) FROM Orders O
INNER JOIN Shippers S
on O.ShipperID = S.ShipperID
where S.ShipperName = "Speedy Express";
```

Answer: 54

2b) What is the last name of the employee with the most orders?

```
SELECT LastName
FROM Orders O JOIN Employees E
ON O.EmployeeID = E.EmployeeID
GROUP BY O.EmployeeID
ORDER BY COUNT(O.OrderID) desc
LIMIT 1;
```

Answer:
Last Name – Peacock
Orders – 40

2c) What product was ordered the most by customers in Germany?

```
SELECT P.ProductName FROM OrderDetails D
JOIN Orders O
ON D.OrderID = O.OrderID
JOIN Customers C
ON O.CustomerID = C.CustomerID
JOIN Products P
ON P.ProductID = D.ProductID
WHERE C.Country = "Germany"
GROUP BY P.ProductID
ORDER BY SUM(Quantity) desc
LIMIT 1;
```

Answer:
Boston Crab Meat