Fall 2021 Data Science Intern Challenge

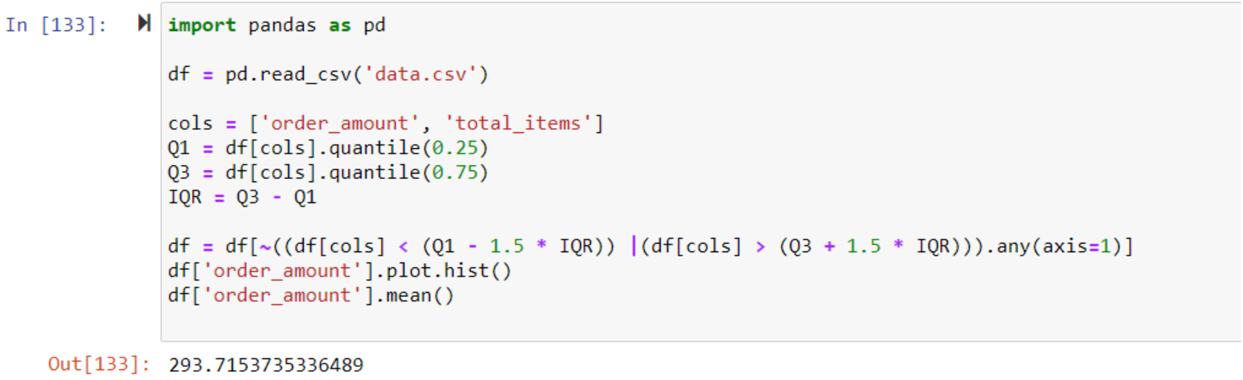
George Saade - gsaade@ryerson.ca

Question 1: Given some sample data, write a program to answer the following: click here to access the required data set

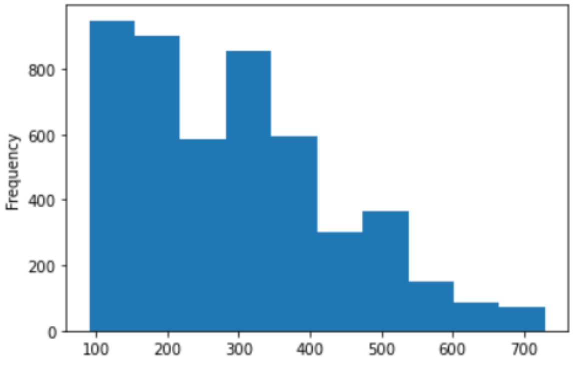
> On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of $3145.13. Given that we know these shops are selling sneakers, a relatively affordable item, something seems wrong with our analysis.

1. **Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.**

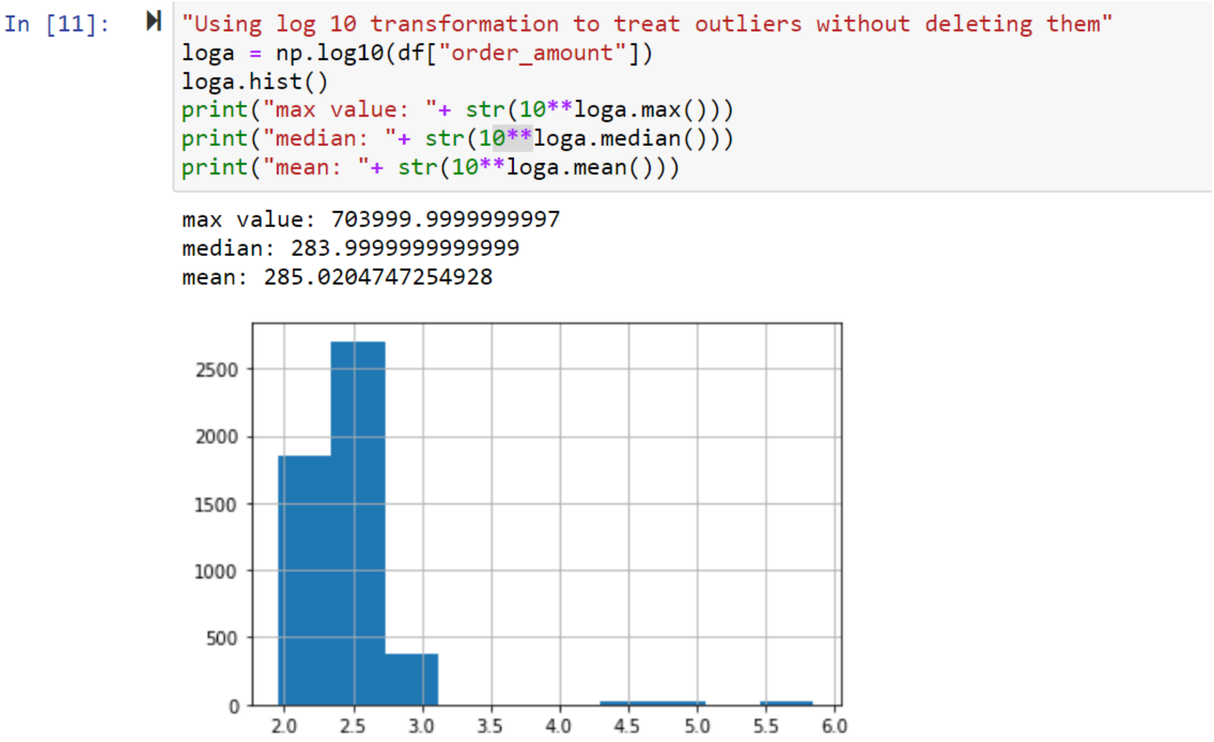
With python:
(explanation on chosen columns provided in the "by hand" section below)



```
In [133]:  import pandas as pd

           df = pd.read_csv('data.csv')

           cols = ['order_amount', 'total_items']
           Q1 = df[cols].quantile(0.25)
           Q3 = df[cols].quantile(0.75)
           IQR = Q3 - Q1

           df = df[~((df[cols] < (Q1 - 1.5 * IQR)) |(df[cols] > (Q3 + 1.5 * IQR))).any(axis=1)]
           df['order_amount'].plot.hist()
           df['order_amount'].mean()

Out[133]:  293.7153735336489
```



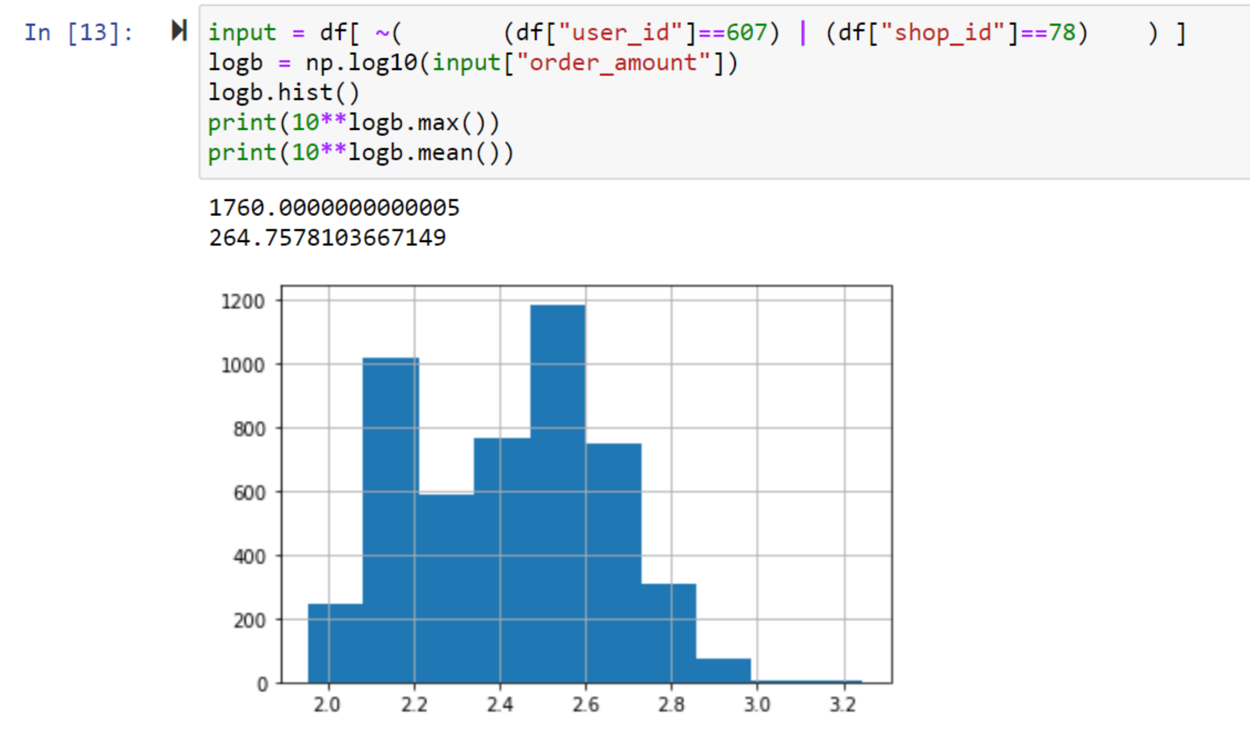
"Using log 10 transformation to treat outliers without deleting them"



```
In [11]:  "Using log 10 transformation to treat outliers without deleting them"
          loga = np.log10(df["order_amount"])
          loga.hist()
          print("max value: "+ str(10**loga.max()))
          print("median: "+ str(10**loga.median()))
          print("mean: "+ str(10**loga.mean()))
```

```
max value: 703999.9999999997
median: 283.9999999999999
mean: 285.0204747254928
```

Log 10 transformation excluding user 607 and shop 78



```
In [13]:  input = df[ ~(       (df["user_id"]==607) | (df["shop_id"]==78)     ) ]
          logb = np.log10(input["order_amount"])
          logb.hist()
          print(10**logb.max())
          print(10**logb.mean())
```

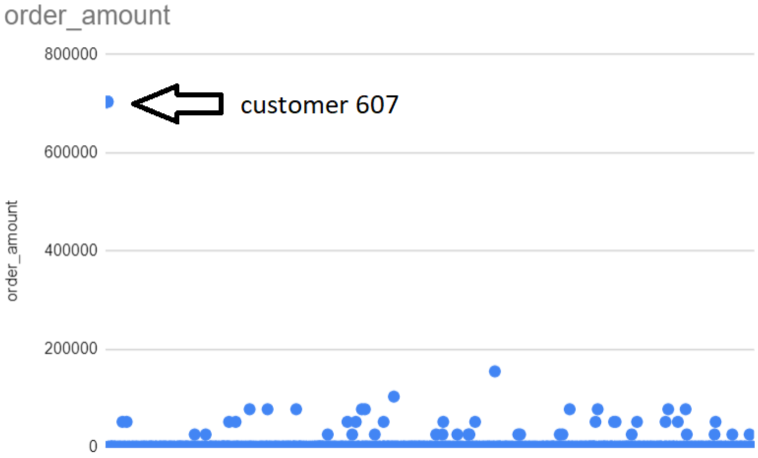
```
1760.0000000000005
264.7578103667149
```

By hand:

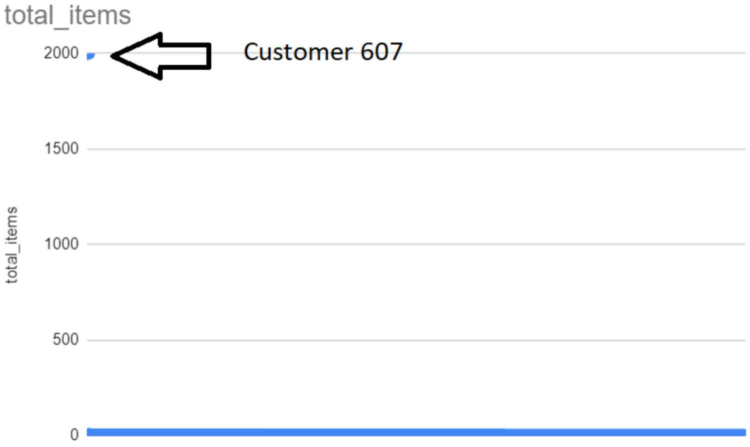
The answer $3145.13 was obtained by calculating the sum of order_amount column values over the total number of orders. 15725640/5000 = $3145.13

The median is $284 which represents a more realistic number we are trying to approach.

Let’s investigate by sorting the order_amount column. One can notice that the biggest purchase is $704000 which is repeated 17 times by customer 607. While the next biggest purchase is only $154350. A difference of 704000-154350 = $549650.

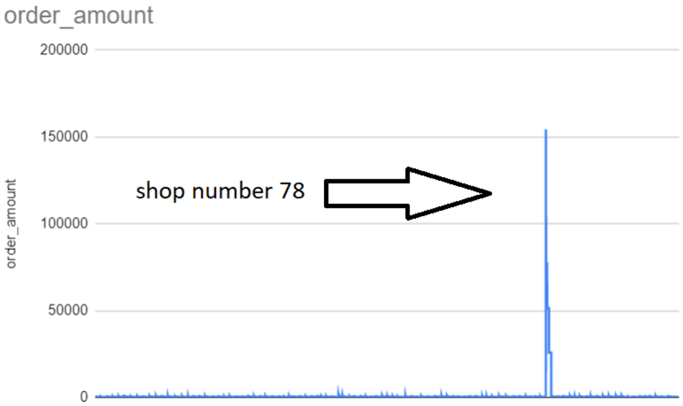


By sorting the data by total items bought we can tell that the greatest total item amount also comes from customer 607 which ordered 2000 items per order. While the next highest number of total ordered items is only 6 items for the order. This is a clear case of an outlier.



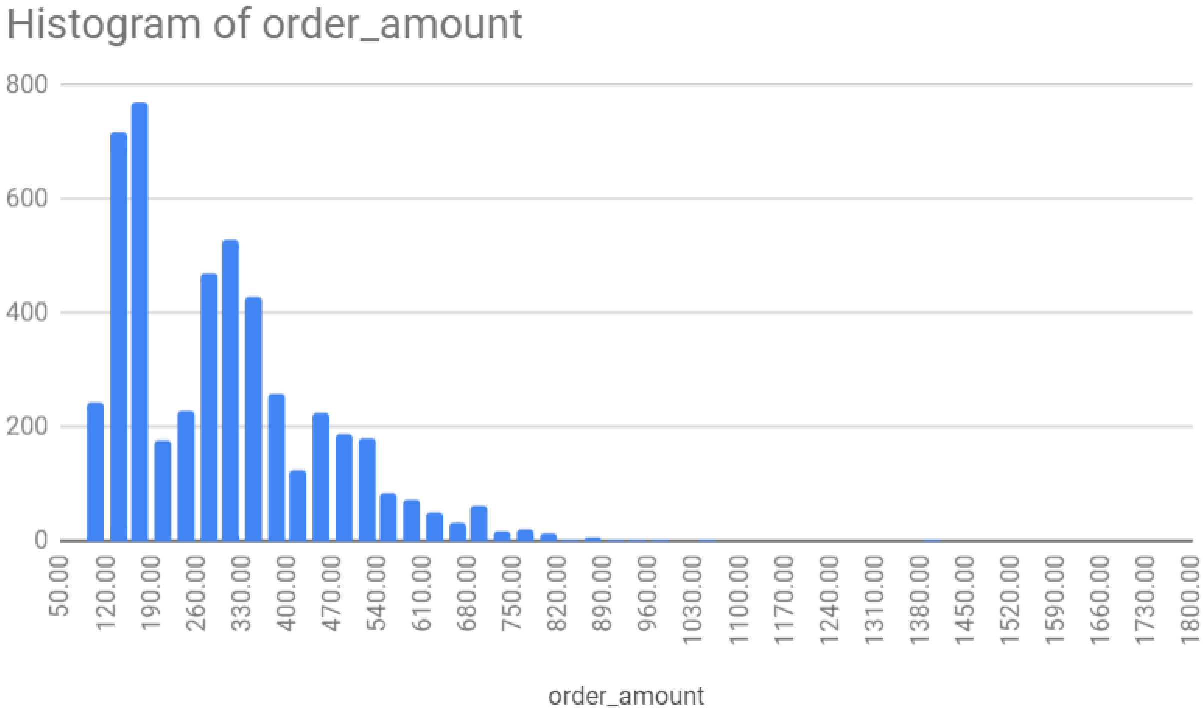
Removing the outlier manually results in AOV of 3757640/4983 = $754.09

Another outlier observed is shop number 78. Its lowest order is $25725. The next largest order amount is $1760 for shop number 42.

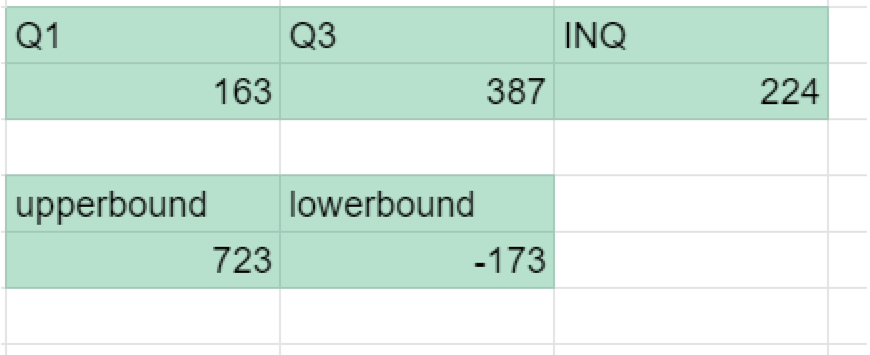


Removing the outlier manually results in AOV of 1519565/4938 = $302.58

We only removed 5000 - 4938 = 62 which is 1.24% of the instances so we should be fine.



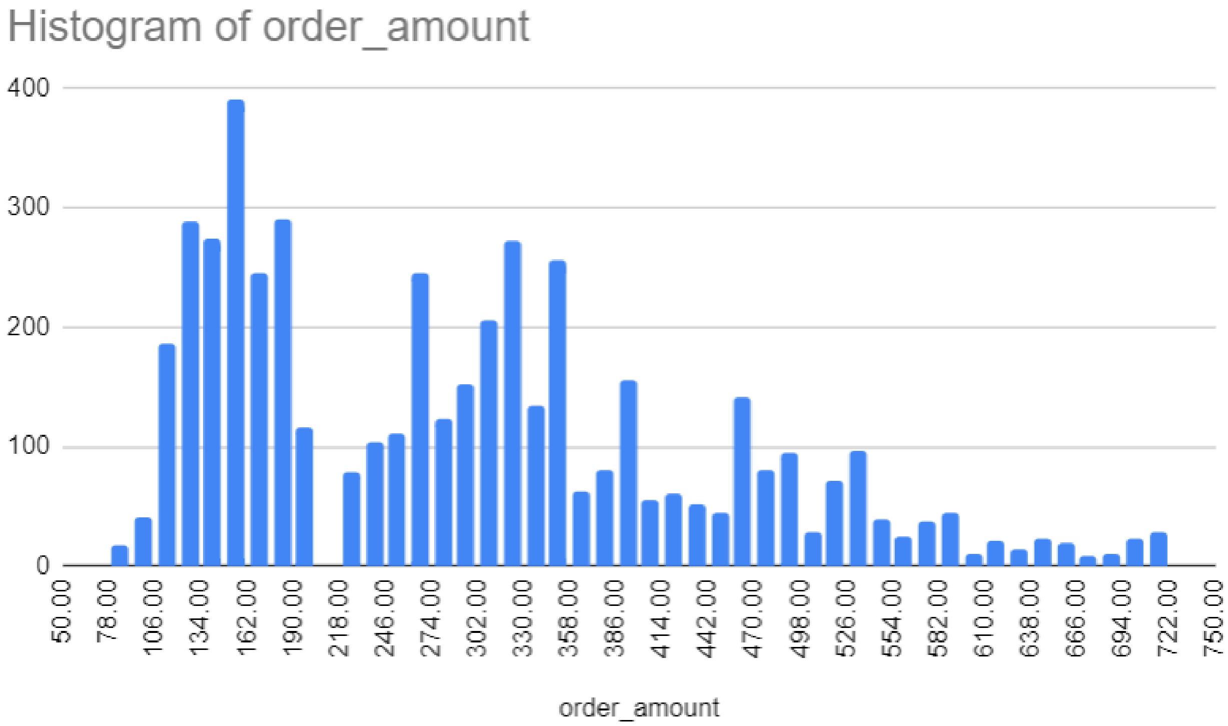
Histogram gives the impression of a lot of outliers around the $820 price mark and above. Now using quartiles to calculate bounds and remove excess outliers.



Q1	Q3	INQ
163	387	224

upperbound	lowerbound
723	-173

New average order value is 1423536/4854= $293.27
With 5000 - 4854 = 146 instances removed in total , which is 2.92% of the data
Median is $280



2. **What metric would you report for this dataset?**

average order value (AOV)

3. **What is its value?**

   $293.27

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?

```
SELECT Count(*)
FROM Orders
INNER JOIN Shippers ON Orders.ShipperID=Shippers.ShipperID AND
Shippers.Shippername="Speedy Express" ;
```

54 records

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

```
SELECT LastName
FROM(
SELECT COUNT(*) COUNTT, Employees.LastName FROM Orders
LEFT JOIN Employees ON Orders.EmployeeID=Employees.EmployeeID
Group BY Orders.EmployeeID
ORDER BY COUNTT DESC  LIMIT 1)
```

Peacock

c. What product was ordered the most by customers in Germany?
(there can be more efficient ways but i preferred readability)

```
SELECT ProductName
FROM(
SELECT ProductName ,SUM(Quantity) Orders
FROM OrderDetails
LEFT JOIN Products ON Products.ProductID=OrderDetails.ProductID
WHERE OrderID IN (
SELECT Orders.OrderID FROM Orders
LEFT JOIN Customers ON Orders.CustomerID=Customers.CustomerID
WHERE Customers.Country="Germany")
GROUP BY OrderDetails.ProductID
ORDER BY Orders DESC  LIMIT 1)
```

```
Boston Crab Meat
```