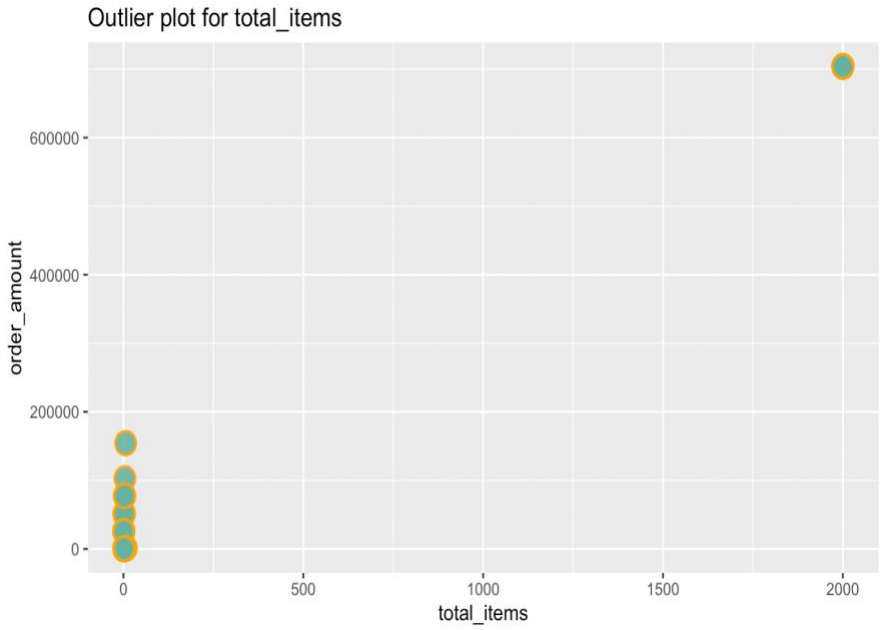
Fall 2022 Data Science Intern Challenge

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.
>
> Ans) To begin with, sneakers are relatively affordable However few metrics or observations from the data has been misleading the Average Order value.
>
> Proposed Steps for Analysis:
> Step1: Exploratory Data Analyses to understand the data
> Step2; Identify the outlier's presence with Visualization.
> Step3: Clean the data, Apply Data transformation and summarize.
> Step4: Conclude the Average order value of sneakers.

a. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.
   Visualizing the order amount and total_items with scatter plots has shown some outliers in the plots which are majorly affecting the AOV.



With summarizing and Group by operations on the data, **it is Observed that Shop_id 78 has an AverageOrderValue of $25725**, which is way more than other shop AOV.

A tibble: 100 × 2

shop_id <dbl>	averagebyshop <dbl>
78	25725
42	352
12	201
89	196
99	195
50	193
38	190
6	187
51	187
11	184

Figure 1 AOV By Shop_id

A tibble: 8 × 2

total_items <dbl>	Number_of_occurances_In_DF <int>
2000	17
8	1
6	9
5	77
4	293
3	941
2	1832
1	1830

Figure 2 total_items Occurrences.

The above table shows that there are 2000 Items Ordered from some particular shop where there is a possibility that it can be wrong order registered.

A tibble: 17 × 2

shop_id <dbl>	total_items <dbl>
42	2000
42	2000
42	2000
42	2000
42	2000
42	2000

Figure 3 Shop_id of total_items=2000

From theses table it is observed that the all the order with toal_items 2000 are from the same shop_id.

b. What metric would you report for this dataset?

```r
summary(cleaned_shopify_df$order_amount)
mean(cleaned_shopify_df$order_amount)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   90.0   163.0   284.0   302.6   387.0  1760.0
[1] 302.5805
```

The mean average is $302.6 after cleaning the data, but still look like way costlier than normal sneakers cost, it has the min of $90 but the max range is $1760

which means the data distribution might have some data which is not normally distributed, skewing the results.

It has the min

It is more relevant to calculate the Average order value by Average Per Order. Average per Order = Order_amount/total_items. And analyze the mode, median and mean metrics comparatively to choose better metric.

```r
describe(cleaned_shopify_df$Amount_per_Order)
```

```
cleaned_shopify_df$Amount_per_Order
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90
    4937        0       57    0.999    151.8    29.53      112      117      132      153      166      181
     .95
     190

lowest :  90  94 101 111 112, highest: 193 195 196 201 352
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   90.0   132.0   153.0   151.8   166.0   352.0
[1] "Mode Value"
[1] 153
[1] "meadian Value"
[1] 153
[1] "Mean Value"
[1] 151.7885
```

After certain data cleaning, data manipulation and aggregation techniques the cleaned data has mode: 153, mean value: 151.78, median value of : 153. Thus, for these particular kind of data all these three metric might work as they are almost near values.

As we removed some data, from metric perspective it can be said that mean average of 302.6 can be improved by the **median and mode value**.

c. What is its value?
   The average order value of the sneaker would be approximately $152 as per the analysis.

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?
   Ans)54
   SELECT Shippers.ShipperName, COUNT(Orders.OrderID) AS TotalOrdersShipped
   FROM Orders

JOIN Shippers ON Orders.ShipperID = Shippers.ShipperID where ShipperName = 'Speedy Express'
GROUP BY ShipperName;

Figure 4 Output by ShipperName

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

SELECT Employees.LastName, Employees.EmployeeID, count(Orders.OrderID) as TotalOrders from Employees join Orders on Employees.EmployeeID = Orders.EmployeeID group by Orders.EmployeeID order by TotalOrders desc limit 1;

Figure 5 Output by Most Orders

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

```
select p.ProductName as MostOrdered_product, o.OrderID, c.country 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
```

Figure 6 Product Order most by Germany people