Fall 2022 Data Science Intern Challenge

Q1:

import openpyxl

# loading the worksheet

wb = openpyxl.load\_workbook("2019 Winter Data Science Intern Challenge Data Set.xlsx")

# get the data set

ws = wb.worksheets[0]

# the way how questioner calculate AOV

data = ws.iter\_rows(min\_row=2, max\_row=5001, min\_col=4, max\_col=5)

amount0 = 0

for row in data:

amount0 += row[0].value

amount0 = amount0 /5000

# Exclude the extreme value in 'total\_item', and delete the value more than two digits in 'total\_item', which is distinguished by '10' here

data = ws.iter\_rows(min\_row=2, max\_row=5001, min\_col=4, max\_col=5)

items = []

for row in data:

if row[1].value > 10:

items.append(row[1].value)

print("The amount of considerable orders is "+str(len(items)))

def solution():

# The AOV should be:

data = ws.iter\_rows(min\_row=2, max\_row=5001, min\_col=4, max\_col=5)

total\_amount = 0

total\_order = 0

total\_amount\_con = 0

total\_order\_con= 0

for row in data:

if row[1].value <= 10:

total\_amount += row[0].value

total\_order += 1

#The average of considerable orders

else:

total\_amount\_con += row[0].value

total\_order\_con += 1

aov = round(total\_amount / total\_order,2)

aov\_con = total\_amount\_con / total\_order\_con

ret = [aov, aov\_con]

return ret

print("AOV should be "+str(solution()[0])+".\nBesides, AOV of considerable orders is "+str(solution()[1])+".")

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

Answer:Exclude the extreme value in 'total\_item', and delete the value more than two digits in 'total\_item', which is distinguished by '10' here.

b.What metric would you report for this dataset?

Answer：I will choose to use the median to reflect the real order situation. The robustness and validity of statistics and the computational complexity in practical application are the most important measures in data statistics. As far as robustness is concerned, the median is obviously better. The common indicator to measure robustness is the collapse point, that is, the maximum proportion that can "distort" statistics. For the mean, only one value is infinite, and the mean will be infinite. But to change the median to infinity, you need to increase the value of half the data, so the median is much more robust than the mean. The data in the worksheet yields 17 sets of configurable data that distort the average, so I think the median is a better representation.

c.What is its value?

AOV should be $754.09.

Q2:

1. How many orders were shipped by Speedy Express in total?

SELECT count(OrderID)

FROM Orders O,Shippers S

where S.ShipperID=O.ShipperID

and S.shippername='Speedy Express'

Result: 54

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

SELECT LastName

FROM Orders o , Employees e

where o.EmployeeID = e.EmployeeID

GROUP BY o.EmployeeID

ORDER BY COUNT(\*) DESC

Result:Peacock

1. What product was ordered the most by customers in Germany?

SELECT ProductName

FROM Orders o,Customers c,OrderDetails od, Products p

Where o.CustomerID = c.CustomerID

and o.OrderID = od.OrderID

and p.ProductID = od.ProductID

and Country='Germany'

ORDER BY Quantity DESC

Result:Steeleye Stout