SALIH ILASLAN   salih_ilaslan@yahoo.com

Winter 2021 Data Science Intern Challenge Answers

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.

ANSWER 1: *I used Python and Jupyter Notebook to answer this questions and I attached my Jupyter file including code and answers to github link. Below, I also gave answer to given questions.*

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

There could be outliers in order value part, which may have caused such a difference in the AOV value. In order to determine what's wrong and understand the data better, I would do an exploratory analysis to analyze data set and employ data visualization is necessary. I used Python and Jupyter Notebook to evaluate this data. Sorting the data would be a simple but effective way to highlight unusual values. Therefore, I would sort the amount column and total items column, then look for unusual high values.

b.   What metric would you report for this dataset?
Based on my exploratory analysis, I would report order_amount column and maybe total_items column since there were outlier values in them. These outliers could be resulted from a mistake or wrong entry. And I would check how many outliers are there in the these columns, and recommend to remove outlier values from the data if there are negligible amount of outliers. I also would employ Z-Score method to detect outliers as a supportive analysis..

c.   What is its value?
AOV : $302.58

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.

Answer 2: *Below, I pasted queries and answers for the questions.*

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

SELECT COUNT (OrderID) FROM Orders
JOIN Shippers ON Orders.ShipperID = Shippers.ShipperID
WHERE Shippers.ShipperNAME = 'Speedy Express';

COUNT (OrderID)
54

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

SELECT Employees.LastName, COUNT (OrderID) FROM Orders
JOIN Employees ON Orders.EmployeeID = Employees.EmployeeID
GROUP BY LastName ORDER BY COUNT (OrderID) DESC LIMIT 1;

LastName	COUNT (OrderID)
Peacock	40

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

SELECT ProductName, COUNT (OrderDetails.OrderID) FROM OrderDetails
JOIN Products ON OrderDetails.ProductID = Products.ProductID
JOIN Orders ON OrderDetails.OrderID = Orders.OrderID
JOIN Customers ON Orders.CustomerID = Customers.CustomerID
WHERE Customers.Country = 'Germany' GROUP BY Products.ProductName
ORDER BY COUNT (OrderDetails.OrderID) DESC LIMIT 1;

ProductName	COUNT (OrderDetails.OrderID)
Gorgonzola Telino	5