# Winter 2021 Data Science Intern Challenge Solutions

Name: Vikram Dhingra

Email: [vikram.dhingra@queensu.ca](mailto:vikram.dhingra@queensu.ca)

Github: <https://github.com/vikramdhingra/shopify_ds_challenge>

**Question 1:** Given some sample data, write a program to answer the following: [click here to access the required data set](https://docs.google.com/spreadsheets/d/16i38oonuX1y1g7C_UAmiK9GkY7cS-64DfiDMNiR41LM/edit#gid=0)

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.

Solution: https://github.com/vikramdhingra/shopify\_ds\_challenge/blob/master/Q1.R

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

* We will calculate summary statistics to check for any outliers
* Visualizing data for spotting outliers/ trends in the data
* Manipulating the data for rectified AOV

1. What metric would you report for this dataset?

* Weekly Total Sales
* Weekly Average Sale
* Week Over Week Growth
* Total sales and order amount for every Shop

1. What is its value?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** | **Sales\_Total** | **Sales\_avg** | **Week\_diff** | **WoW** |
| **9** | 2598496 | 3760.486252 | 0 | 0.0% |
| **10** | 2699341 | 2256.97408 | -1503.512172 | -66.6% |
| **11** | 3351952 | 2984.819234 | 727.8451539 | 24.4% |
| **12** | 4195635 | 3686.849736 | 702.0305022 | 19.0% |
| **13** | 2880216 | 3380.535211 | -306.3145251 | -9.1% |

Below is the snippet for 3 of 100 shops:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **shop\_id** | **Sales\_total\_amount** | **Sales\_avg\_order\_amnt** | **Sales\_total\_vol** | **Sales\_avg\_order\_vol** |
| 42 | 11990176 | 235101.4902 | 34063 | 667.9019608 |
| 78 | 2263800 | 49213.04348 | 88 | 1.913043478 |
| 50 | 17756 | 403.5454545 | 92 | 2.090909091 |

**Question 2:** For this question you’ll need to use SQL. [Follow this link](https://www.w3schools.com/SQL/TRYSQL.ASP?FILENAME=TRYSQL_SELECT_ALL) 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.

Solution: https://github.com/vikramdhingra/shopify\_ds\_challenge/blob/master/Q2.sql

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

-- A1

SELECT count(distinct(OrderID))

FROM [Orders] A LEFT JOIN [Shippers] B ON A.ShipperID=B.ShipperID

WHERE ShipperName = 'Speedy Express'

-- 54

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

-- A2

SELECT B.LastName,

count(distinct(OrderID)) AS total\_order

FROM [Orders] A LEFT JOIN [Employees] B ON A.EmployeeID=B.EmployeeID

GROUP BY A.EmployeeID

ORDER BY total\_order DESC

LIMIT 1

-- LastName total\_order

-- Peacock 40

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

There is ambiguity with regards to the product orders. There can two scenarios:

1. **We sum the Quantity and group the product for Germany orders. If we follow this approach; below is the query and result**

-- A3

SELECT C.ProductName,

C.ProductID,

COUNT(A.OrderID) AS total\_order,

SUM(B.Quantity) AS Product\_quant

FROM [Orders] A LEFT JOIN [OrderDetails] B ON A.OrderID=B.OrderID -- 196 and 518

LEFT JOIN [Products] C ON B.ProductID=C.ProductID

LEFT JOIN [Customers] D ON A.CustomerID=D.CustomerID

WHERE D.Country ='Germany'

GROUP BY C.ProductName,

C.ProductID

ORDER BY Product\_quant DESC

LIMIT 1

-- ProductName ProductID total\_order Product\_quant

-- Boston Crab Meat 40 4 160

1. **We count order and group the the product for Germany orders. If we follow this approach; below is the query and result**

SELECT C.ProductName,

C.ProductID,

COUNT(A.OrderID) AS total\_order,

SUM(B.Quantity) AS Product\_quant

FROM [Orders] A LEFT JOIN [OrderDetails] B ON A.OrderID=B.OrderID -- 196 and 518

LEFT JOIN [Products] C ON B.ProductID=C.ProductID

LEFT JOIN [Customers] D ON A.CustomerID=D.CustomerID

WHERE D.Country ='Germany'

GROUP BY C.ProductName,

C.ProductID

ORDER BY total\_order DESC

LIMIT 1

-- ProductName ProductID total\_order Product\_quant

-- Gorgonzola Telino 31 5 125