Data Science Challenge – Fall 2022

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.

A. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.
B. What metric would you report for this dataset?
C. What is its value?

Answers:

1. The AOV calculation done in the question results in an abnormally large value of $3145.13. Most likely, the AOV was calculated by using the following equation:

$$AOV = \frac{Total\ Order\ Amount}{Total\ Item\ Count}$$

   However, the error would have occurred by calculating the total number of items using the count() function instead of the sum() function. This would result in an error as the count() function only provides the total count of the number of rows containing an integer value in the column ‘total_items’. This does not accurately depict the total amount of items as rows containing total_item values > 1, are still registered as 1 count using the count() function.

   This could be better evaluated using the sum() function to determine the total item count, as that would actually take the sum of the values contained within each cell of the ‘total_items’ row and would result in a more accurate total item value, resulting in a more accurate AOV. However, despite that, the dataset could be further improved by looking for outliers. The average sneaker cost ranges from $70 to $250 [citation]. For the purposes of this question, it would be appropriate to calculate the average cost of each item in each data entry, and remove all data entries who’s average cost of each item is greater than $350.

2. For this dataset, the reporting metrics are the sums of both the ‘order_amount’ and the ‘total_items’.
   order_amount_inc = data['order_amount'].sum()
   total_item_inc = data['order_amount'].count()

   AOV_inc = order_amount_inc / total_item_inc

3. The resultant value of the calculation is an AOV of $150.40 instead of the previous incorrect value of $3,145.13. The boxplot before the dataset was modified to remove the entries with an average cost/item being > $350 was produced:

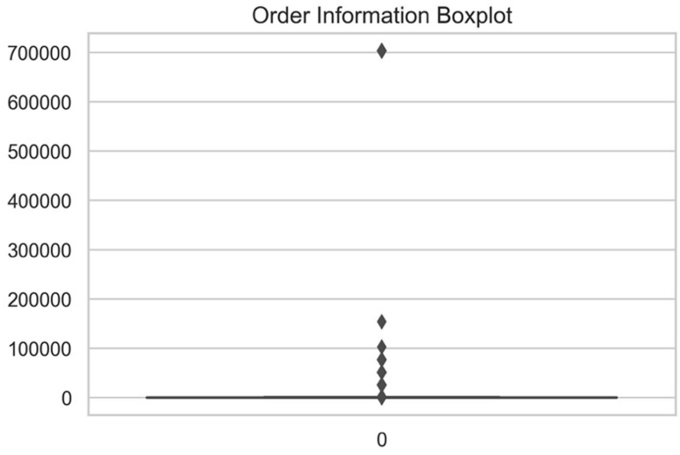

Figure 1: Order Information Boxplot Pre-Processing

The boxplot after the dataset was modified to remove outliers can be seen down below.

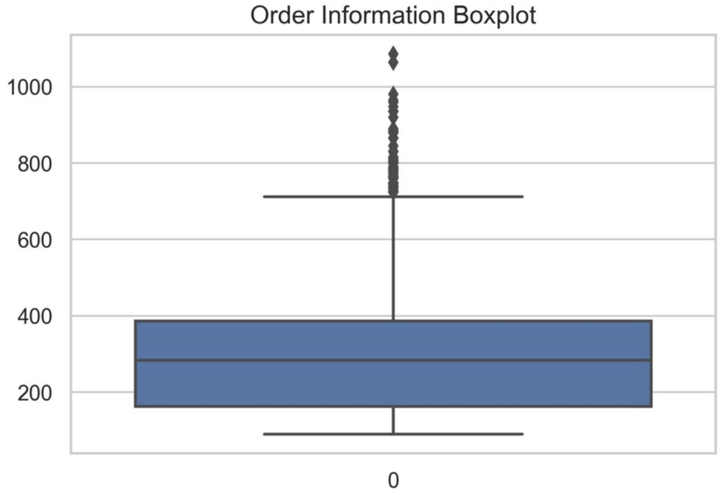

Figure 2: Order Information Boxplot Post-Processing

See Appendix A for Code OR View Code Here: https://github.com/shaheerbee/Shopify-Data-Science-Intern-Application-2022/blob/main/Shaheer_Hassan_Shopify_DataScience_Challenge_Q1.ipynb

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?
b. What is the last name of the employee with the most orders?
c. What product was ordered the most by customers in Germany?

Answers

a. The total number of orders shipped by Speedy Express was 54 orders. (See Appendix B for Code)
b. The last name of the employee with the most orders is Peacock. (See Appendix B for Code)
c. The product that was ordered the most by customers in Germany was Boston Crab Meat. Bost Crab Meat had a total of 160 orders. (See Appendix B for Code)

See Appendix B for code or visit this link: https://github.com/shaheerbee/Shopify-Data-Science-Intern-Application-2022/blob/main/DataScience_Intern_Question2.sql

Appendix A – Question 1 Python

```
#Import Libraries

import os
import csv
from pathlib import Path
import seaborn as sns
import matplotlib.pyplot as plt

#Import the Required Dependencies
import pandas as pd
import numpy as np


# ## Importing the Dataset

#collect the data

data = pd.read_csv('Q1_data.csv')
#data.head()
data


# ## Look at Boxplot for Outliers

sns.set()
sns.set_theme(style="whitegrid")
ax = sns.boxplot(data = data['order_amount'])
plt.title('Order Information Boxplot', size=13)
plt.savefig('Q1_BoxPlot.jpg', dpi=300, format='jpg', bbox_inches = 'tight',
transparent='false')
plt.ylabel('Order Amount ($)')
plt.show()
df= data[data['order_amount'] > 10000]
df


# ## First Average Order Value (AOV) Calculation

order_amount_sum = data['order_amount'].sum()
total_items_sum = data['total_items'].sum()

AOV = order_amount_sum / total_items_sum
print('The Average Order Value is:', '${:,.2f}'.format(AOV))
```

```
# ## Remove Outliers & Re-calculate AOV

#since average sneaker prices range from $70 to 250$, we will calculate the ratio
between total order cost
# and total items to find the average cost of each item.
#To be safe, if the average cost of the item in a specific data entry > $300, we will
remove it

df_new = data['order_amount'] / data['total_items']

#add it to the exisiting df

data['item_cost_avg'] = df_new
data

#remove all rows in which item_cost_avg > $300

df_final = data[data['item_cost_avg'] < 350]
df_final

order_amount_final = df_final['order_amount'].sum()
total_items_final = df_final['total_items'].sum()

AOV_final = order_amount_final / total_items_final
print('The Average Order Value is:', '${:,.2f}'.format(AOV_final))

sns.set()
sns.set_theme(style="whitegrid")
ax = sns.boxplot(data = df_final['order_amount'])
plt.title('Order Information Boxplot', size=13)
plt.savefig('Q1_BoxPlot_2.jpg', dpi=300, format='jpg', bbox_inches = 'tight',
transparent='false')
plt.ylabel('Order Amount ($)')
plt.show()


# ## Incorrect AOV Calculation from Question
#
# * This calculation is done to make sure that the AOV value provided in the question
was calculated via the following incorrect process

order_amount_inc = data['order_amount'].sum()
total_item_inc = data['order_amount'].count()

AOV_inc = order_amount_inc / total_item_inc
print('The Incorrect Average Order Value is:', '${:,.2f}'.format(AOV_inc))
```

Appendix B – Question 2 SQL

```sql
/* Question 1 */

SELECT COUNT(*) as total_shipment_speedexpress
FROM Shippers as s
LEFT JOIN Orders as o
ON s.ShipperID = o.ShipperID
Where ShipperName = "Speedy Express"

/* Question 2 */

FROM Employees as e
LEFT JOIN Orders as o
ON e.EmployeeID = o.EmployeeID
GROUP BY e.EmployeeID
ORDER BY num_of_orders DESC
LIMIT 1

/* Question 3 */

SELECT p.ProductName, SUM(Quantity) AS TotalQuantity
FROM Orders AS o, OrderDetails AS od, Customers AS c, Products AS p
WHERE c.Country = "Germany" AND od.OrderID = o.OrderID AND od.ProductID = p.ProductID
AND c.CustomerID = o.CustomerID
GROUP BY p.ProductID
ORDER BY TotalQuantity DESC
LIMIT 1;
```