Summer 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.

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

Answers

- 1. The AOV is influenced by outliers and skewed data. Here you are getting an AOV of \$3145.13 because you are simply calculating the mean of all order amounts. Shop IDs 42 and 78 are the outliers as their order amounts lie in the range [352, 70400] and [25725, 154350] respectively. The range for other shops is much lesser.
- 2. A better way to evaluate this data would be to use the mode instead. Mode is not influenced by outliers. Thus mode is a better measure for tracking the revenue earned per order for businesses. As this is the most frequent order executed by customers. If the modal values are low, implementing strategies to increase the mode will have a positive impact on the revenue.
- 3. The mode for all 100 shops is \$153

Additionally I am a bit confused about the need for finding mode or AOV across all 100 shops. This makes sense only if all 100 shops fall under the same parent organization/owner. If these are separate entities then we should calculate the above metrics specific to each shop data. I have done the same in the notebook below. A lot more insight can be derived from this approach.

Also I see that one order can constitute more than one pair of sneakers. Hence dividing the order amount by the total_items will get us the price of sneakers at each shop. We can further use this new feature in analysing the data and how tis influences the AOV per shop. But I don't think this lies within the scope of this challenge.

```
In [ ]:
         import pandas as pd
         order_file = "./data.csv"
         df = pd.read_csv(order_file, parse_dates=True, infer_datetime_format=True)
         df['created_at'] = pd.to_datetime(df['created_at']).dt.date
Out[]:
           order_id shop_id user_id order_amount total_items payment_method created_at
                                                                  cash 2017-03-13
        0
                       53
                             746
                                         224
        1
                2
                       92
                             925
                                          90
                                                                  cash 2017-03-03
                                                                  cash 2017-03-14
        2
                3
                       44
                             861
                                          144
                                                     1
        3
                                                             credit_card 2017-03-26
                4
                       18
                             935
                                          156
                                                     1
                5
                       18
                             883
                                          156
                                                             credit_card 2017-03-01
                                                     1
         naive_ans = df['order_amount'].mean()
                                                      #Get the AOV computed by you
         naive_ans
Out[]: 3145.128
In [ ]:
         aov\_shop = {}
                                                   # Compute shop specific metrics
         mode_shop = {}
         median shop = {}
         for shop_id in df.shop_id.unique():
             temp_df = df[df['shop_id'] == shop_id]
             aov_shop[shop_id] = temp_df['order_amount'].mean()
             mode shop[shop id] = temp df['order amount'].mode()[0]
             median_shop[shop_id] = temp_df['order_amount'].median()
         shop df = pd.DataFrame.from dict(aov shop, orient='index', columns=['aov shop'])
         shop df['mode shop'] = pd.DataFrame.from dict(mode shop, orient='index', columns=['mode shop'])['mode shop']
         shop df['median shop'] = pd.DataFrame.from dict(median shop, orient='index', columns=['median shop'])['median shop']
         shop_df.head()
```

aov_shop mode_shop median_shop

Out[]:

1 of 2

		aov_shop	mode_shop	median_shop
	53	214.117647	224	224.0
	92	162.857143	180	180.0
	44	262.153846	144	288.0
	18	342.588235	156	312.0
		054040453	130	2700
n []:		erall_mode		der_amount'
ut[]:	153			
in []:				

2 of 2