**Background**

The consumer team is fortunate enough to have multiple actionable items within the retail shoe space. However, the team is getting both credit card data and overall company financials such as revenue, margins, EBITDA, and free cash flow from the investment bankers. Looking for a competitive edge against other private equity bidders, you turn to email receipt data which can identify exact itemized shoe SKUs (stock-keeping units). Surprisingly (and for the sake of the exercise), all eight shoe companies for sale are the exact same price of $1 billion dollars in order to keep things simple. The ultimate question posed to you by the managing partners and investment committee is **which shoe company Advent should purchase**? Additionally, you will need to answer the below structured questions in to support your decision. Note that you will depend **solely** on the provided US email receipt data for this deal, and no other external information or dataset will be a factor.

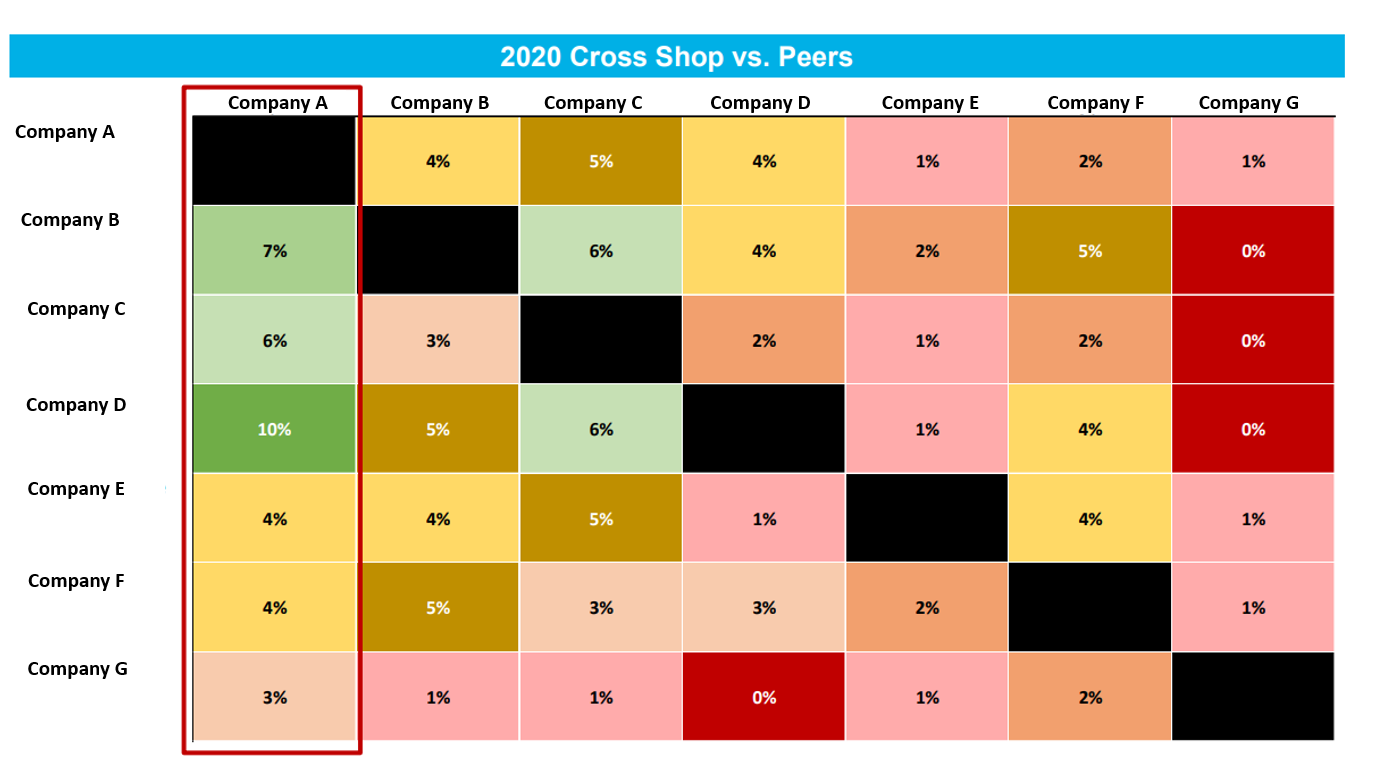
**Instructions**

Write your code in **Python** but show the output in Excel. You may type out a word document or PowerPoint presentation if you feel that Excel is an unwieldy tool to explain your answer to the above question. Please submit all relevant files prior to the deadline. Note that we only want shoe purchases and not apparel or other accessories even if they are from the below companies (i.e. exclude Puma socks and Skechers t-shirts) and also ignore the corporate parent (i.e. Uggs is part of Deckers).

* Allbirds
* Crocs
* Dr. Martens
* Puma
* Skechers
* Steve Madden
* Timberland
* UGG

**Specific Questions to Answer:**

1. Pull together the email receipt data by quarters instead of its current format of days starting from **1Q19 to 4Q20**. 1Q19 is defined as January 1, 2019 to March 31, 2019. 2Q19 is defined as April 1, 2019 to June 30, 2019, and so on. (**Note: please use “email\_time” as order date**)
2. Calculate the number of unique buyers of each of the 8 shoe companies.
   1. Create an 8x8 matrix and calculate how many unique customers from each company also bought from each of the other 7 companies. Please see below as an example of this matrix output below except in % form:



1. Please calculate quarterly retention rate for each of the 8 companies
   1. Hint: For 4Q20, this is defined as the % of unique customers from 3Q20 (July 1, 2020 to September 30, 2020) that also bought at **least once** in 4Q20 (October 1, 2020 to December 31, 2020)
2. For each of the 8 shoe companies, what % of unique customers in full year 2019 and 2020 **only buy** from that shoe brand?
3. For each of the 8 companies, how many different shoe SKUs are there in total?
   1. Also provide the name of the most popular shoe SKU (make sure you clearly define why that item is a separate shoe type)
   2. Using the above most popular shoe item across the 8 companies, provide the average pricing of the shoe SKU over the eight quarters from 1Q19 to 4Q20. Which specific shoe has seen the most pricing increase over the past 2 years?
4. Force-rank the 8 shoe companies by the $ value of their customers by multiplying the below 2 data points together.
   1. Calculate the average order size for each shoe company (i.e. $135.42)
   2. Calculate the average frequency of orders in any given quarter (i.e. 2.1 purchases a quarter)
5. Which shoe company would you purchase and why? Be as creative and analytical as possible and support your answer with the email receipt data. Make sure you clearly show your logic and workflow!

**DATA DICTIONARY**

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Notes |
| merchant\_name | name of vendor | an exhaustive list of vendors is available |
| user\_id | Hashed unique email address identifier | This field links receipt data with the user level dataset |
| order\_number | order number included in the receipt email | Default value is empty. |
| order\_time | Timestamp of the purchase, if included in the receipt denoted as email recipient’s local time zone. | Default value is empty. |
| email\_time | UTC timestamp the email arrived in a user’s inbox | The date an email was received may not necessarily be the date of purchase. |
| insert\_time | UTC timestamp indicating when the receipt was first captured in the Edison database | Ideal case here would be for insert\_time to be close to email\_time. Two cases in which this is not the case: (a) historical receipts sent before a user joined the panel and (b) receipts sent before Edison was tracking a particular vendor or particular receipt. |
| update\_time | UTC timestamp when this particular receipt values were modified | Default value is empty. Updates happen as part of an effort to add more fields to the extraction or correct an improper extraction. |
| email\_subject | subject line for the transaction’s email | Names and contact information are redacted to safeguard privacy. |
| order\_total\_amount | total cost for the order | Default value is empty. |
| order\_subtotal | total cost of all the items purchased | Default value is empty. |
| order\_points | number of rewards points (e.g. American AAdvantage Miles) were used to complete the purchase | Default value is empty. When this field is not empty, order\_total\_amount might be misleading if cash was used only to pay for tax or as a partial payment in conjunction with rewards points. |
| order\_shipping | shipping cost for the order | Default value is empty. |
| order\_tax | total taxes for the order | Default value is empty. |
| order\_discount | total discount, if any, listed on the order | Default value is empty. |
| order\_currency | Currency symbol in the receipt if present | Default value is empty. |
| from\_domain | address of the sender email address (e.g. ‘mail.acmecorp.com’ for receipts sent from ‘noreply@mail.acmecorp.com’) | 0 |
| product\_description | text describing the product/item included in the receipt | Default value is empty. |
| product\_subtitle | text further describing the product/item included in the receipt, if relevant | Default value is empty. |
| item\_quantity | number ordered for each item | Default value is ‘1’. |
| order\_total\_qty | total number of items in the order | Default value is ‘1’. |
| item\_price | price for the item | Default value is empty. |
| digital\_transaction | boolean value indicating whether the purchase was for a digital good (e.g. ebook, software subscription) |  |
| SKU | store keeping unit value, if present in the receipt | Default value is empty. |
| item\_id | unique value identifying the particular item, if in the receipt | Default value is empty. |
| checksum | unique identifier for the emailed receipt | Functionally, identifies unique orders. Reference this field when writing us to investigate an issue with the extraction. |
| order\_pickup | 0 = online purchase delivered to home, 1 = online purchase picked up at store, 2 = email confirmation for in-store purchase, 3 = return, 4 = purchase email for seller | Default value is ‘0’. Seller, in this case, can represent a marketplace seller receiving confirmation that their item has sold (e.g. a store on eBay) or a restaurant receiving confirmation of a takeout order. |
| order\_discount | total discount, if any, listed on the order | Default value is empty. |
| product\_reseller | the proximate seller of the item | Default value is empty. Usually empty except for marketplaces where the reseller is explicitly listed in the receipt |
| delivery\_date | date the record was delivered; corresponds with file name the record is included in UTC | Date format: YYYY-MM-DD (Ex: 2019-01-24) |
|  |  |  |