E-Commerce Dataset

Context

The provided dataset has information on 100k orders from 2016 to 2018 from a large e-commerce market place. Its features allow you to view an order from multiple dimensions : order status, product attributes, price, etc. The merchants are able to sell their products through the marketplace and ship them directly to the customers.

Assignment

As a consultant at Sia Partners you are given the assignment to analyze these datasets and ***conduct a customer segmentation***. Use the available data, your data skills and your business sense. Your solution should demonstrate both consultancy and data science skills. We ask you to spend between 4 to 8 hours preparing the case. This is a relatively short time, so do not spend all your time on data analysis and coding, but also take the time to put together a good business presentation.

Your presentation should include the following aspects:

1. Sales trend analysis over time
2. Customer analysis
3. Customer segmentation based on the data provided.
4. **Optional – Further Exploratory analysis**: What other interesting aspects (not mentioned above) did you find while analyzing these datasets? Do not hesitate to create additional features to illustrate your findings.
5. Conclusion with main insights

Requirements

- Use the provided data as a basis for your analysis and approach.
- There is no coding language imposed, but you will be required to share your code with us.
- Make a PowerPoint presentation to illustrate your findings.

Evaluation Criteria

- Clear storyline
- Interpretability of data visuals
- Applicability of your analysis (including customer segmentation)
- Creativity
- Tidy code

Interview

During the interview, we will ask you to present your findings and customer segmentation results, as you would to the client. Think carefully about what is relevant to your audience. After the presentation, we will walk through your code together in detail and ask technical questions about it.

Code

Send your code the day before the interview, so that we have a chance to evaluate it before we start the interview. Make a zip and send it directly to sang.ahn@sia-partners.com

Enclosed Datasets

- **customers_dataset.csv**
- **orders_dataset.csv**
- **order_items_dataset.csv**
- **products_dataset.csv**
- **product_category_name.csv**

Description of datasets

1. **customers_dataset.csv** : This dataset groups the customers information (customer id and geolocation information). There are duplicate customer_unique_id, because the dataset is per order. Where client_id is the key to the order dataset. Each order has a unique client_id.

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
0	06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
1	18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	9790	sao bernardo do campo	SP
2	4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2b5e	1151	sao paulo	SP

2. **orders_dataset.csv** : This dataset groups the order information (customer, order status, tracking in time), identified by a unique order_id and attached to a customer by the customer_id.

order_id	customer_id	order_status	order_purchase_time	order_approved_at	order_delivered_car	order_delivered_cus	order_estimated_del
e481f51cbdc54678b	9ef432eb625129730	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10 21:25:13	2017-10-18 0:00:00
53cdb2fc8bc7dce0b	b0830fb4747a6c6d2	delivered	2018-07-24 20:41:37	2018-07-26 3:24:27	2018-07-26 14:31:00	2018-08-07 15:27:45	2018-08-13 0:00:00
47770eb9100c2d0c4	41ce2a54c0b03bf34	delivered	2018-08-08 8:38:49	2018-08-08 8:55:23	2018-08-08 13:50:00	2018-08-17 18:06:29	2018-09-04 0:00:00

3. **order_items_dataset.csv :** This dataset groups order information related to each product purchased (product id, quantity = order_item_id, price of product, etc.)

order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value
00010242fe8c5a6d1	1	4244733e06e7ecb4	48436dade18ac8b2	2017-09-19 9:45:35	58.9	13.29
00018f77f2f0320c55	1	e5f2d52b802189ee6	dd7ddc04e1b6c2c6	2017-05-03 11:05:13	239.9	19.93
000229ec398224ef6	1	c777355d18b72b67	5b51032eddd242ad	2018-01-18 14:48:30	199	17.87
00024acbcdf0a6daa	1	7634da152a4610f15	9d7a1d34a5052409	2018-08-15 10:10:18	12.99	12.79
00042b26cf59d7ce6	1	ac6c3623068f30de0	df560393f3a51e745	2017-02-13 13:57:51	199.9	18.14

4. **products_dataset.csv :** This dataset gathers the product categories as well as various characteristics of dimensions and weight of each product.

product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm
1e9e8ef04db	perfumaria	40	287	1	225	16	10	14
3aa071139cb	artes	44	276	1	1000	30	18	20
96bd76ec881	esporte_lazer	46	250	1	154	18	9	15
cef67bcfe19(	bebes	27	261	1	371	26	4	26

5. **product_category_name.csv :** Translated product category in English.

product_category_name	product_category_name_english
beleza_saude	health_beauty
informatica_acessorios	computers_accessories
automotivo	auto

Data Schema