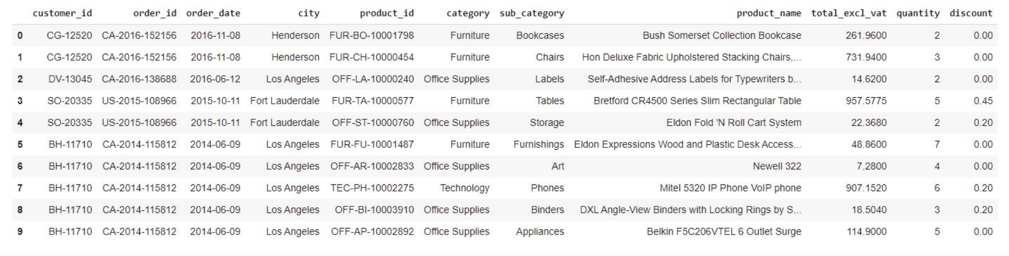
Data scientist candidate task

Robert Tarasevič

July 21, 2021

1 Task and data overview

For the customer segmentation task, sales data were provided (figure 1).



	customer_id	order_id	order_date	city	product_id	category	sub_category	product_name	total_excl_vat	quantity	discount
0	CG-12520	CA-2016-152156	2016-11-08	Henderson	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.9600	2	0.00
1	CG-12520	CA-2016-152156	2016-11-08	Henderson	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs,...	731.9400	3	0.00
2	DV-13045	CA-2016-138688	2016-06-12	Los Angeles	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters b...	14.6200	2	0.00
3	SO-20335	US-2015-108966	2015-10-11	Fort Lauderdale	FUR-TA-10000577	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	957.5775	5	0.45
4	SO-20335	US-2015-108966	2015-10-11	Fort Lauderdale	OFF-ST-10000760	Office Supplies	Storage	Eldon Fold 'N Roll Cart System	22.3680	2	0.20
5	BH-11710	CA-2014-115812	2014-06-09	Los Angeles	FUR-FU-10001487	Furniture	Furnishings	Eldon Expressions Wood and Plastic Desk Access...	48.8600	7	0.00
6	BH-11710	CA-2014-115812	2014-06-09	Los Angeles	OFF-AR-10002833	Office Supplies	Art	Newell 322	7.2800	4	0.00
7	BH-11710	CA-2014-115812	2014-06-09	Los Angeles	TEC-PH-10002275	Technology	Phones	Mitel 5320 IP Phone VoIP phone	907.1520	6	0.20
8	BH-11710	CA-2014-115812	2014-06-09	Los Angeles	OFF-BI-10003910	Office Supplies	Binders	DXL Angle-View Binders with Locking Rings by S...	18.5040	3	0.20
9	BH-11710	CA-2014-115812	2014-06-09	Los Angeles	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C206VTEL 6 Outlet Surge	114.9000	5	0.00

Figure 1: Data preview.

This table was made up of:

1. 793 unique customers.
2. 5009 unique orders.
3. 531 cities.
4. 3 categories.
5. 17 sub categories.
6. 1850 unique products.
7. Sales amount.
8. Discount.
9. Quantity
10. Data was collected from the beginning of 2014 and until the end of 2017

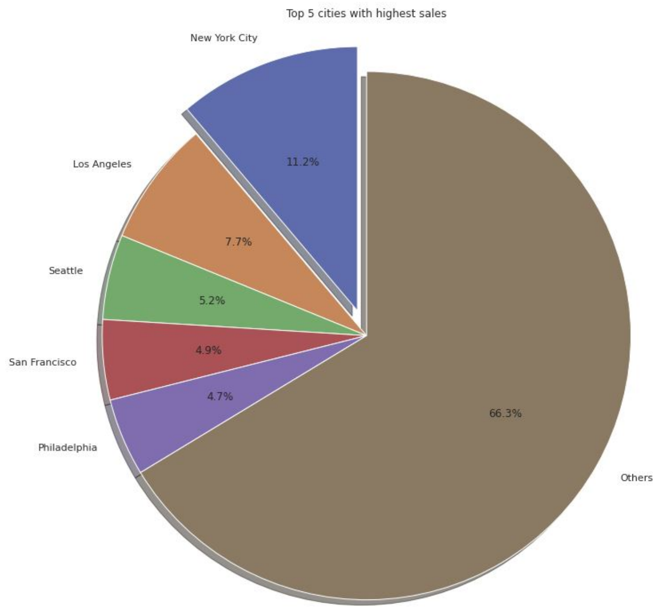


Figure 2: TOP 5 states by sale.

The analysis shows that there are a few states that account for the main part of the revenue, as shown in the pie chart (fig. 2).
Before segmenting all customers, it would be necessary to analyse customer behaviour and habits across states, but this time, for simplicity's sake, I have chosen one state for further segmentation - San Francisco.

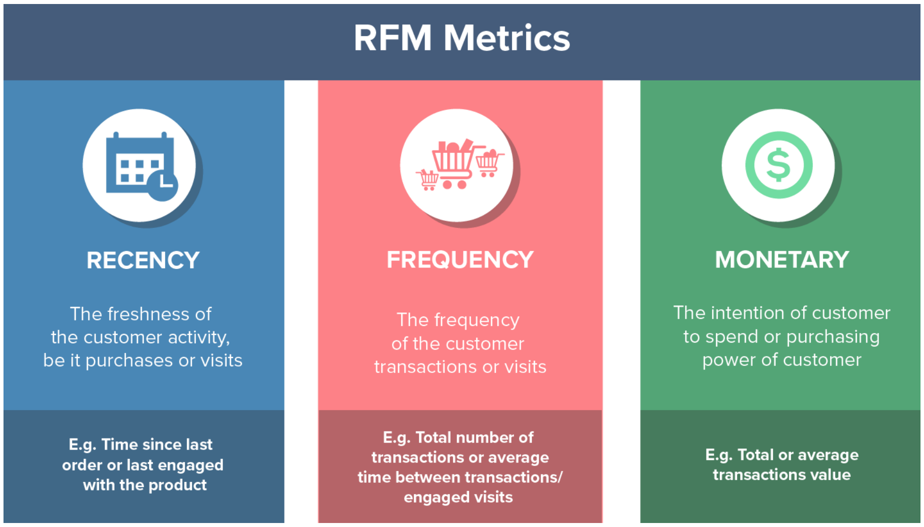
2 San Francisco customers segmentation based on RFM

RFM analysis stands for recency, frequency and monetary. These metrics are important indicators of a customers behaviour and illustrate these facts:

- the more recent the purchase, the more responsive the customer is to promotions;
- the more frequently the customer buys, the more engaged and satisfied they are;
- monetary value differentiates heavy spenders from low-value purchasers

In our case, the RFM analysis and the K-means clustering algorithm identified 3 clusters-segments (based on silhouette score). Looking at these groups in more detail, I can make the following comments, San Francisco State's customers can be divided into 3 groups:

1. ***Lost or partially lost customers*** who were infrequent shoppers but had a large shopping basket and bought from all product groups. These customers were given the biggest discounts but were not attracted, possibly due to a lack of quality or other reasons (further research is needed). Nevertheless, some customers in this group can still be recovered by trying



RFM Metrics

RECENCY

The freshness of the customer activity, be it purchases or visits

E.g. Time since last order or last engaged with the product

FREQUENCY

The frequency of the customer transactions or visits

E.g. Total number of transactions or average time between transactions/ engaged visits

MONETARY

The intention of customer to spend or purchasing power of customer

E.g. Total or average transactions value

Figure 3: RFM.

to contact them, enquiring about the quality of the goods purchased and possibly offering an exclusive offer or introducing new products.

2. The next group could be described as ***one-time buyers***. The majority of customers in this group made only one purchase and their shopping basket was rarely more than $70. These may be buyers who are not attached to a particular seller, but simply ”hunt” for discounts, so it is unlikely that they would be attracted by anything other than a low price. In the event of a sale of cheaper products, contact this group of buyers, they should not miss this opportunity.

3. The last group are the ***very best buyers***, these buyers are frequent purchasers and typically spend around $500 per order. Obviously this group of buyers is the smallest of all and should be given more attention. This group of buyers often buys accessories, binders, paper, less often tables and chairs. So you should maintain a warm relationship with these buyers and from time to time treat them to discounts or small gifts, because it is very likely that they will recommend you to their friends.

Finally, if we look at the overall monthly sales graph, we can see that at the beginning of 2016, sales were slightly down, but in the middle of the year, they started to rise and the upward trend has continued to the present day.

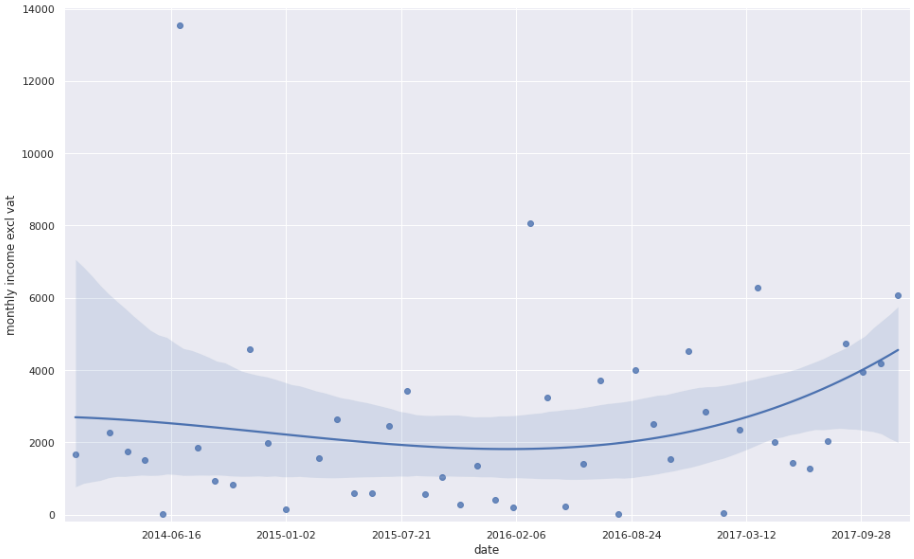


Figure 4: Monthly sales, 3rd order polynomial fit