Machine Learning – Clustering on BigTrade Dataset

By : Ghazali Akmal Rabbani

Content

- Data Preprocessing
- Item that Customer Buy the Most and Item Should be Ignore
- Strategy to Increase Sales based on Our Export Destination Countries
- Customer Segmentation
- Hidden insight from the data

Data Preprocessing

- Check Datatype
- Check Missing Value
- Replace the Missing Value with Median for Numerical Column and Mode for Categorical Column
- Add New Column named as **Sales**
- Parse and Separate the Date and Time

Item that Most Buy by Customer (Top 10)

	Negara	Barang	Jumlah	Harga_Satuan	Sales	Kode_Pelanggan
19361	United Kingdom	WORLD WAR 2 GLIDERS ASSTD DESIGNS	48326.0	158.47	12055.96	7545482.0
17293	United Kingdom	JUMBO BAG RED RETROSPOT	43167.0	5015.27	84516.44	30852990.0
19272	United Kingdom	WHITE HANGING HEART T-LIGHT HOLDER	37689.0	19862.57	131611.47	148833056.0
18121	United Kingdom	POPCORN HOLDER	34365.0	822.58	32425.81	12443019.0
15711	United Kingdom	ASSORTED COLOUR BIRD ORNAMENT	33679.0	2442.80	54662.15	21720588.0
17752	United Kingdom	PACK OF 12 LONDON TISSUES	25307.0	230.27	7639.64	8003697.0
17786	United Kingdom	PACK OF 72 RETROSPOT CAKE CASES	24702.0	942.91	15607.49	18331883.0
19106	United Kingdom	VICTORIAN GLASS HANGING T-LIGHT	23242.0	1719.76	31683.17	15779554.0
16047	United Kingdom	BROCADE RING PURSE	22801.0	258.89	5785.47	3742133.0
15719	United Kingdom	ASSORTED COLOURS SILK FAN	20322.0	538.94	16034.24	7552806.0

Item that should be ignore

	Negara	Barang	Jumlah	Harga_Satuan	Sales	Kode_Pelanggan
8353	Germany	ROUND CAKE TIN VINTAGE GREEN	1.0	0.0	0.0	12647.0
10884	Netherlands	POLKADOT RAIN HAT	144.0	0.0	0.0	14646.0
12700	RSA	Manual	1.0	0.0	0.0	12446.0
15520	United Kingdom	20713	-400.0	0.0	0.0	15152.0
15521	United Kingdom	20713 wrongly marked	-200.0	0.0	0.0	15152.0
15599	United Kingdom	?	-1384.0	0.0	0.0	712144.0
15600	United Kingdom	? sold as sets?	-1200.0	0.0	0.0	15152.0
15601	United Kingdom	??	-1849.0	0.0	0.0	106064.0
15602	United Kingdom	?? missing	-170.0	0.0	0.0	15152.0
15603	United Kingdom	???	-390.0	0.0	0.0	15152.0

These are the item that should be ignore on the dataset. And there are **180 data** which similar to that kind of data

Strategy to Increase Sales

Strategy 1 : Reduce the activity in Top 10 Least Sales by Customer

	Negara	Barang	Jumlah	Harga_Satuan	Sales	Kode_Pelanggan
15640	United Kingdom	AMAZON FEE	-30.0	249042.680	-221520.500	515168.0
17638	United Kingdom	Manual	3663.0	129871.210	-65139.640	7491183.0
16389	United Kingdom	CRUK Commission	-16.0	7933.430	-7933.430	225536.0
16074	United Kingdom	Bank Charges	-13.0	7505.641	-7175.639	558189.0
16575	United Kingdom	Discount	-1191.0	4914.070	-5028.980	1104897.0
18431	United Kingdom	SAMPLES	-59.0	3175.490	-3049.390	954576.0
13484	Spain	Manual	-3.0	2842.620	-2842.620	37463.0
18133	United Kingdom	POSTAGE	-22.0	19055.800	-1408.570	2163775.0
4178	EIRE	Manual	-2.0	15226.450	-1127.130	179864.0
3623	EIRE	Discount	-1.0	434.510	-434.510	14911.0

One of things that can be decrease is activity when using Amazon Fee.
So that we can save more amount Total Sales which deficit on Sales column

Strategy 2 : Do Improvement in Top Most Sales

	Negara	Barang	Jumlah	Harga_Satuan	Sales	Kode_Pelanggan
16547	United Kingdom	DOTCOM POSTAGE	707.0	206252.06	206245.48	10725872.0
18294	United Kingdom	REGENCY CAKESTAND 3 TIER	10376.0	26182.21	134405.94	28438347.0
19272	United Kingdom	WHITE HANGING HEART T-LIGHT HOLDER	37689.0	19862.57	131611.47	148833056.0
17861	United Kingdom	PARTY BUNTING	16709.0	9532.07	92501.73	25246516.0
17293	United Kingdom	JUMBO BAG RED RETROSPOT	43167.0	5015.27	84516.44	30852990.0
17830	United Kingdom	PAPER CHAIN KIT 50'S CHRISTMAS	18197.0	3945.60	61888.19	18136777.0
15711	United Kingdom	ASSORTED COLOUR BIRD ORNAMENT	33679.0	2442.80	54662.15	21720588.0
16230	United Kingdom	CHILLI LIGHTS	10077.0	4519.63	52986.86	10346041.0
17907	United Kingdom	PICNIC BASKET WICKER 60 PIECES	61.0	1299.00	39619.50	30196.0
15877	United Kingdom	BLACK RECORD COVER FRAME	11293.0	1547.92	39387.00	5609589.0

Do improvement in selling on Top 10 Most Sales. Such as DOTCOM POSTAGE.
So the more things like DOTCOM POSTAGE to be sold, the more sales that we got

Strategy 3 : Reconsidering when using Discount

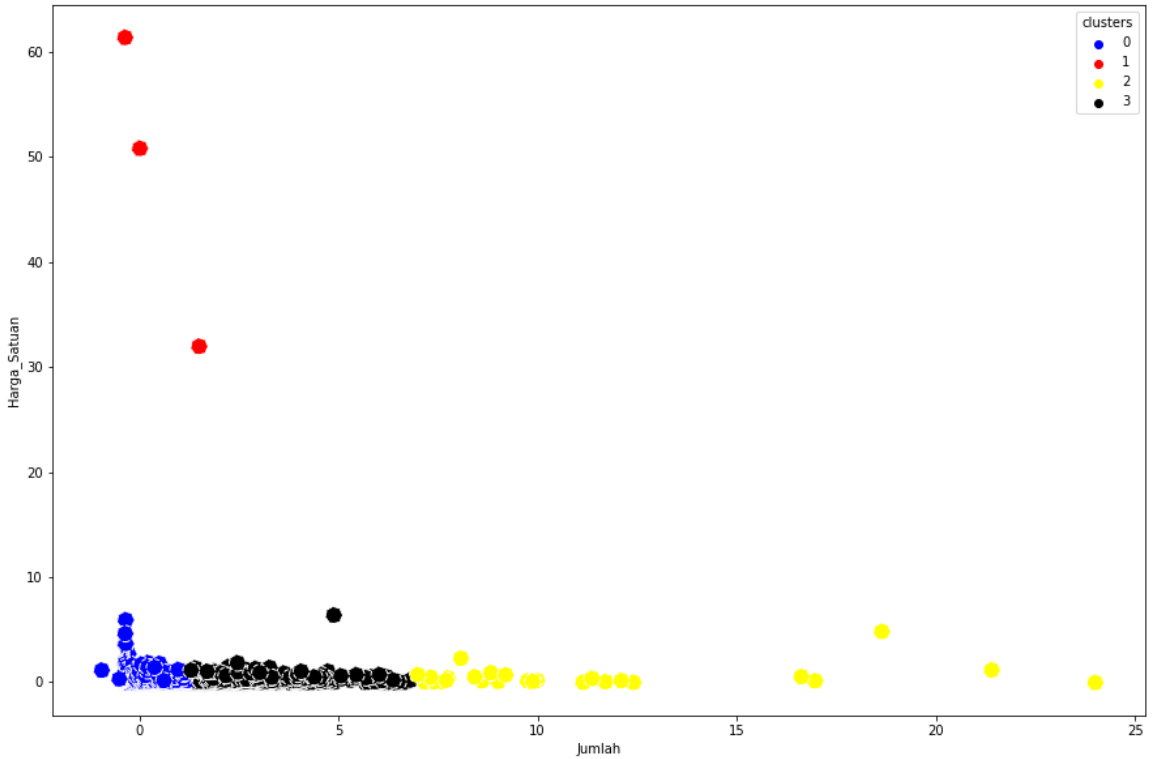
	Barang	Jumlah	Harga_Satuan	Sales	Kode_Pelanggan
0	Discount	-1194.0	5581.31	-5696.22	1149366.0

	Negara	Barang	Jumlah	Harga_Satuan	Sales	Kode_Pelanggan
3623	EIRE	Discount	-1.0	434.51	-434.51	14911.0
9636	Italy	Discount	-1.0	26.33	-26.33	14912.0
10562	Netherlands	Discount	-1.0	206.40	-206.40	14646.0
16575	United Kingdom	Discount	-1191.0	4914.07	-5028.98	1104897.0

Reconsidering when using Discount. Because we already use Discount for customer for 1994 times.
And the dominant of discount that used are in United Kingdom

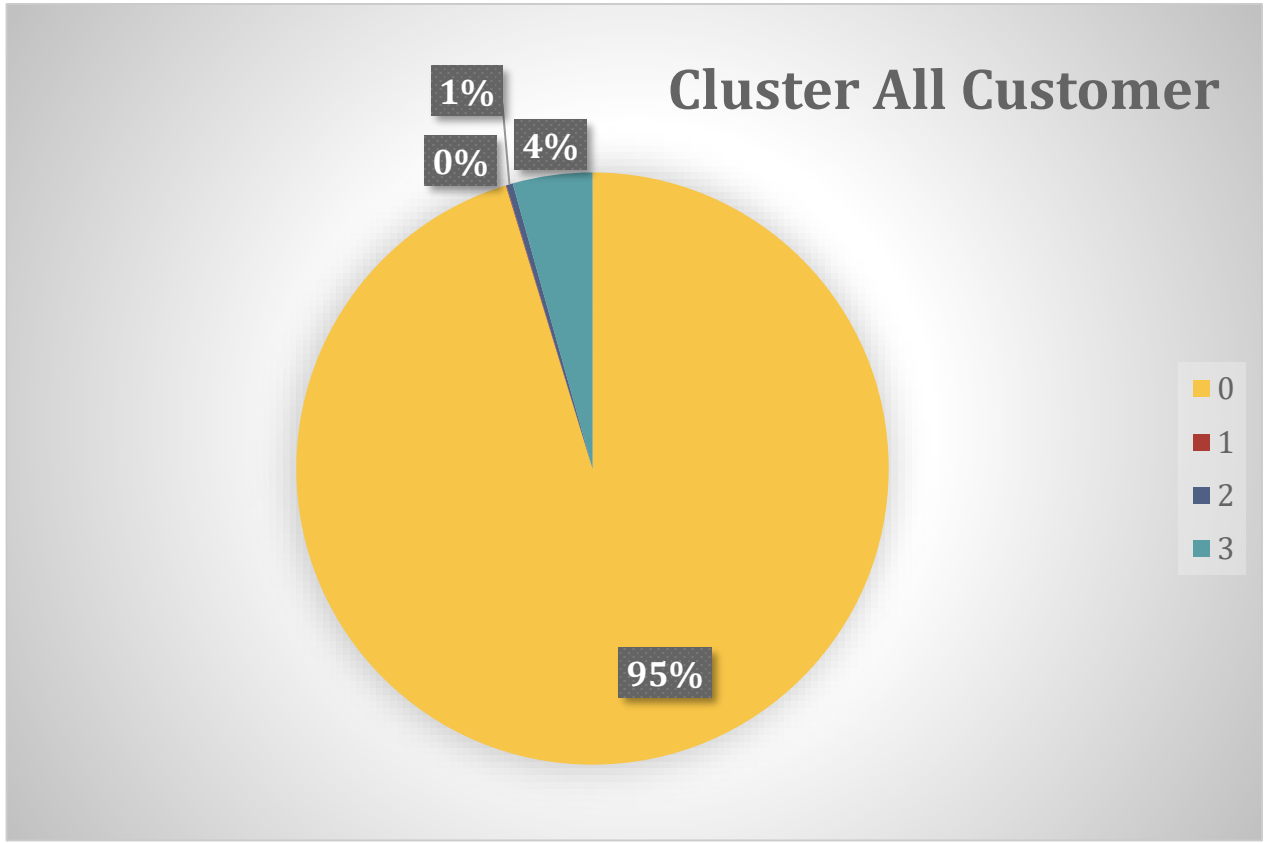
Customer Segmentation

Customer Segmentation Based on All Customer



In this segmentation, I used Jumlah and Harga_Satuan column to figuring out Purchasing Power of Customer. The summary is there are a lot customer who buy the things from range (0 – 10) for Jumlah, and 0 to Below 10 for Harga_Satuan

Percentage of Customer Segmentation Based on All Customer



Cluster	Total Customer
0	7348
1	3
2	27
3	335

Hidden Insight

Statistic of All Customer based on Harga_Satuan

```
==================================
Statistical Data Summary
==================================
Mean :  127.11394861139338
Median :  4.5
1st Quartile :  1.66
3rd Quartile :  17.115000000000002
IQR :  15.455000000000002
Upper Fence :  40.29750000000001
Lower Fence :  -21.52250000000004
==================================
```

Conclusion :

Based on Statistic, Average purchasing of customer based on Harga_Satuan is 127

Total of Middle-up Class based on Harga_Satuan

	Negara	Barang	Jumlah	Harga_Satuan	Sales	Kode_Pelanggan
369	Australia	POSTAGE	0.0	612.73	87.27	24830.0
804	Austria	POSTAGE	37.0	548.00	1456.00	174219.0
1397	Belgium	POSTAGE	272.0	1503.00	4248.00	1217358.0
1442	Belgium	REGENCY CAKESTAND 3 TIER	47.0	191.25	599.25	187088.0
1819	Canada	POSTAGE	1.0	550.94	550.94	17444.0
...	...	...	...	...	...	...
19447	United Kingdom	ZINC T-LIGHT HOLDER STAR LARGE	1809.0	171.27	1733.59	2405798.0
19449	United Kingdom	ZINC T-LIGHT HOLDER STARS SMALL	4409.0	248.22	3680.95	3986796.0
19450	United Kingdom	ZINC TOP 2 DOOR WOODEN SHELF	8.0	285.03	193.33	209479.0
19451	United Kingdom	ZINC WILLIE WINKIE CANDLE STICK	2287.0	279.01	2212.67	3865731.0
19452	United Kingdom	ZINC WIRE KITCHEN ORGANISER	27.0	138.07	208.77	217013.0

2076 rows × 6 columns

Conclusion :

There are 2076 Customers who categorized as Middle-Up Class based on Harga_Satuan

This condition is when harga_satuan more than 127 (Average of Harga_Satuan)

Total of Lower Middle Class based on Harga_Satuan

	Negara	Barang	Jumlah	Harga_Satuan	Sales	Kode_Pelanggan
0	Australia	DOLLY GIRL BEAKER	200.0	1.08	216.0	12415.0
1	Australia	I LOVE LONDON MINI BACKPACK	4.0	4.15	16.6	12393.0
2	Australia	10 COLOUR SPACEBOY PEN	48.0	0.85	40.8	12415.0
3	Australia	12 PENCIL SMALL TUBE WOODLAND	384.0	0.55	211.2	12415.0
4	Australia	12 PENCILS TALL TUBE POSY	252.0	1.14	79.8	24849.0
...	...	...	...	...	...	...
19908	Unspecified	WRAP POPPIES DESIGN	25.0	0.42	10.5	16320.0
19909	Unspecified	WRAP SUKI AND FRIENDS	25.0	0.42	10.5	16320.0
19910	Unspecified	WRAP VINTAGE PETALS DESIGN	25.0	0.42	10.5	16320.0
19911	Unspecified	WRAP WEDDING DAY	25.0	0.42	10.5	14265.0
19912	Unspecified	ZINC METAL HEART DECORATION	2.0	1.25	2.5	12743.0

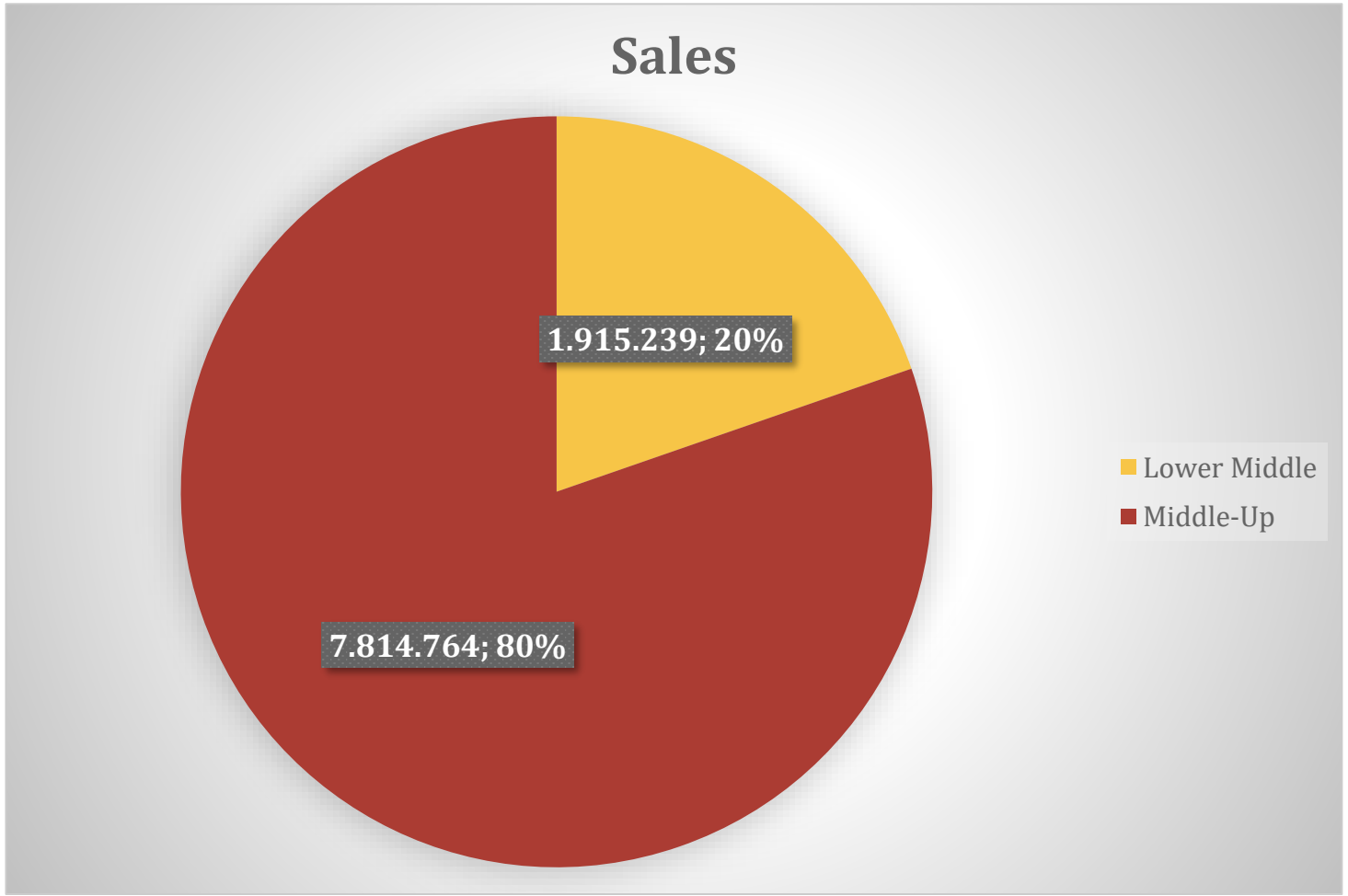
17656 rows × 6 columns

Conclusion :

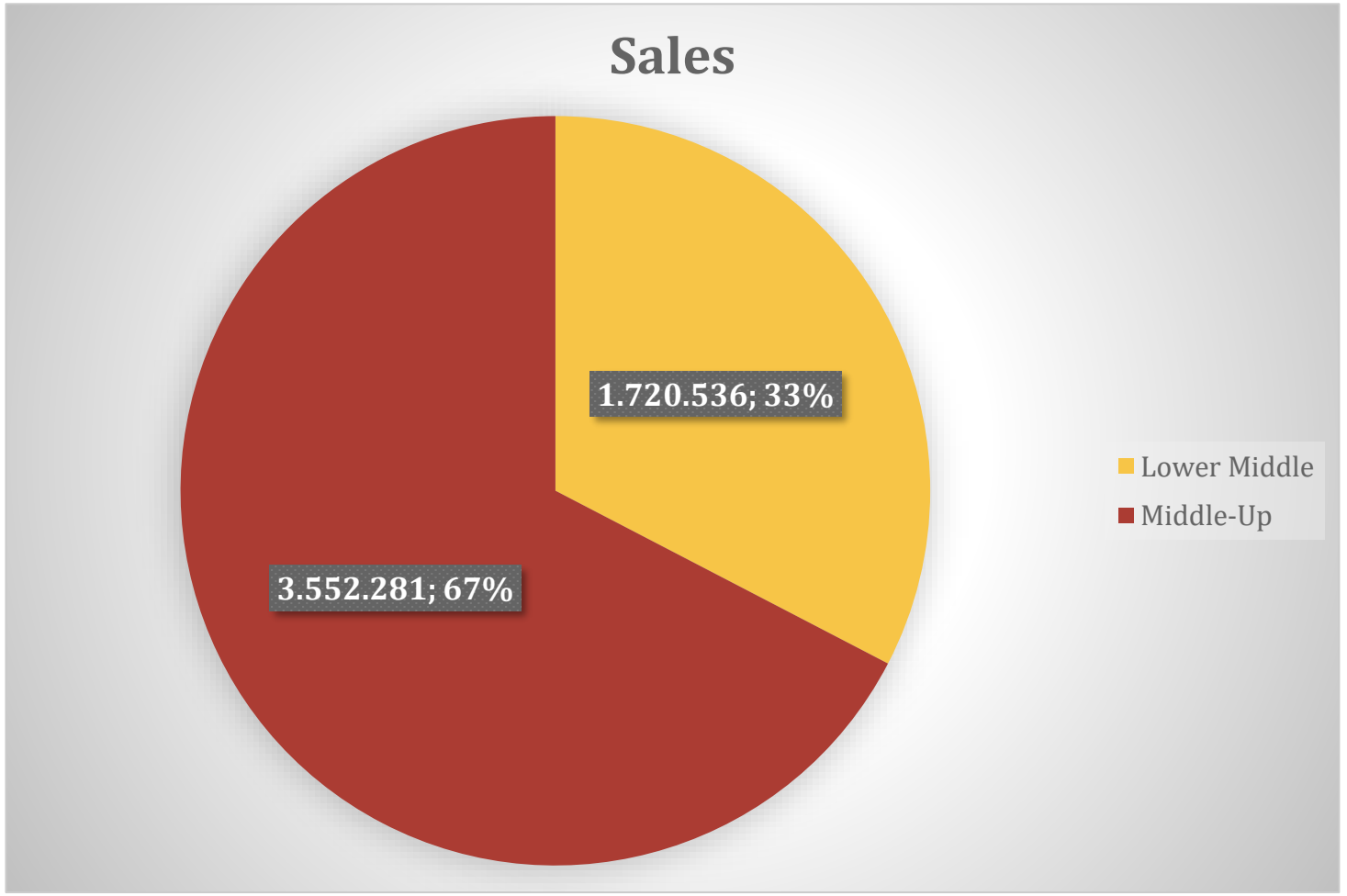
There are 17656 Customers who categorized as Lower Middle Class based on Harga_Satuan

This condition is when harga_satuan lower than 127 (Average of Harga_Satuan)

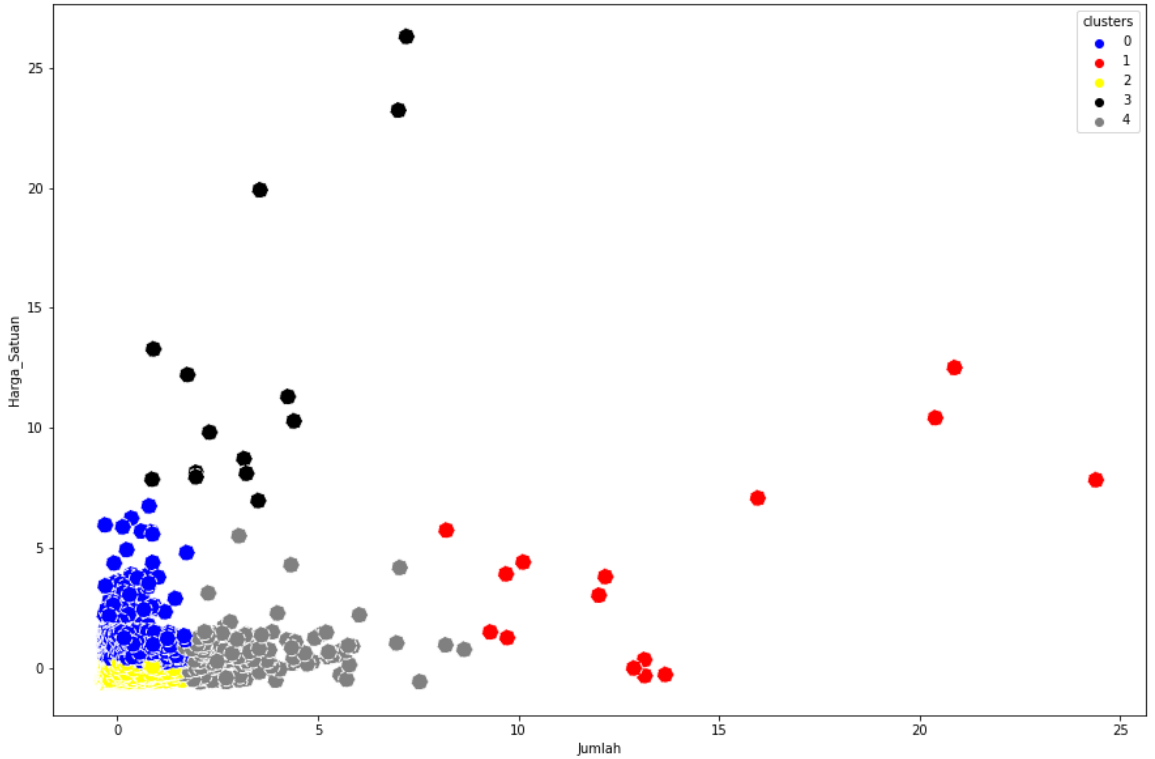
Ratio LowerMiddle vs Middle-Up Contribution based on Sales



Ratio LowerMiddle vs Middle-Up Contribution based on Jumlah

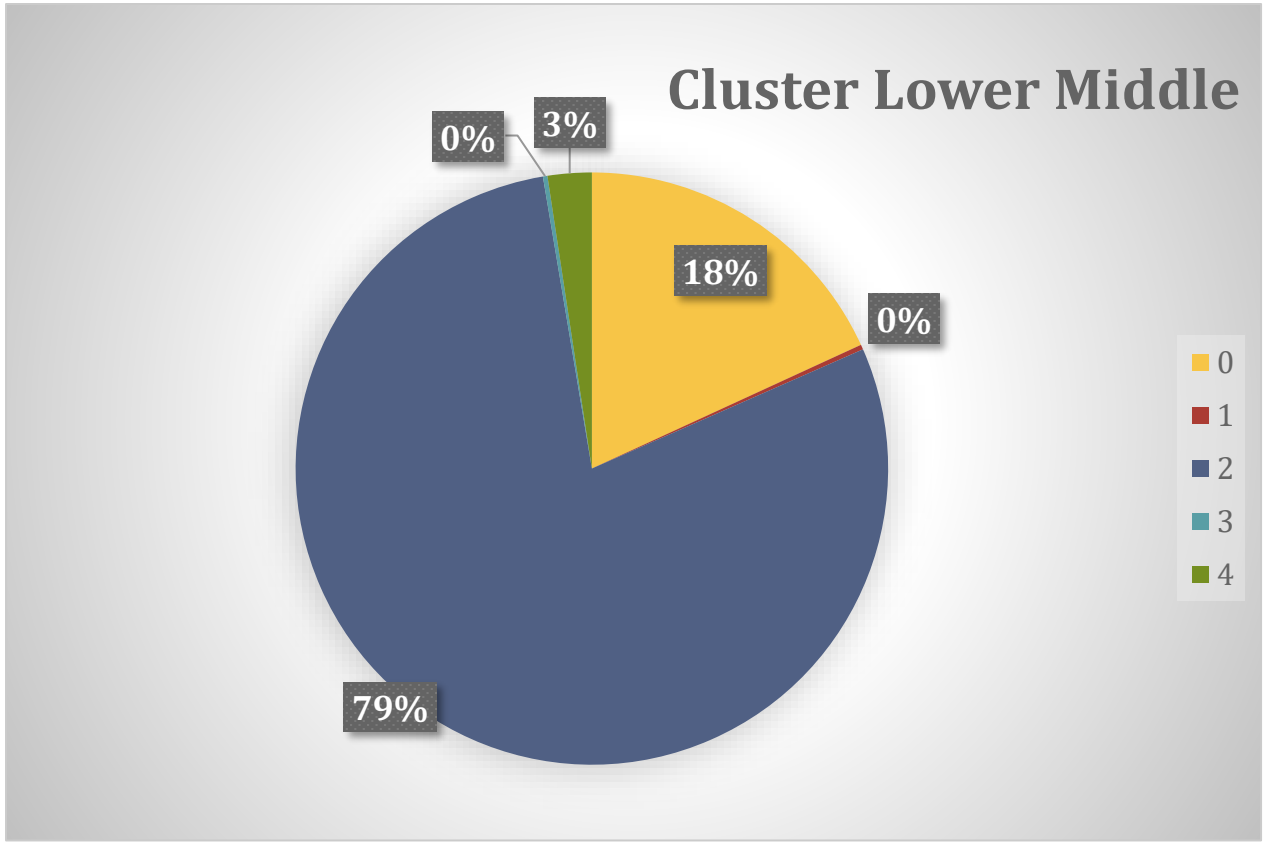


Customer Segmentation Based on Lower Middle Class



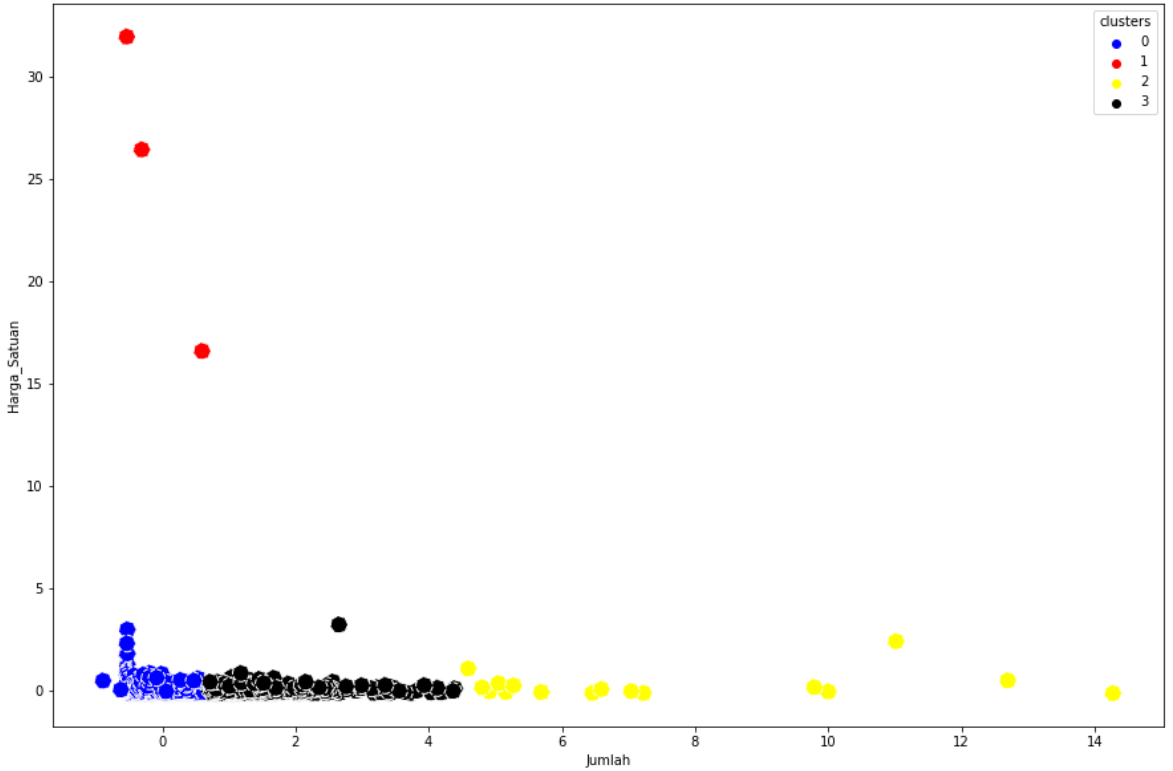
In this segmentation, I used Jumlah and Harga_Satuan column to figuring out Purchasing Power of Customer. The summary is there are a lot customer who buy the things from range (0 – 6) for Jumlah, and 0 to 6 for Harga_Satuan

Percentage of Customer Segmentation Based on Lower Middle Customer



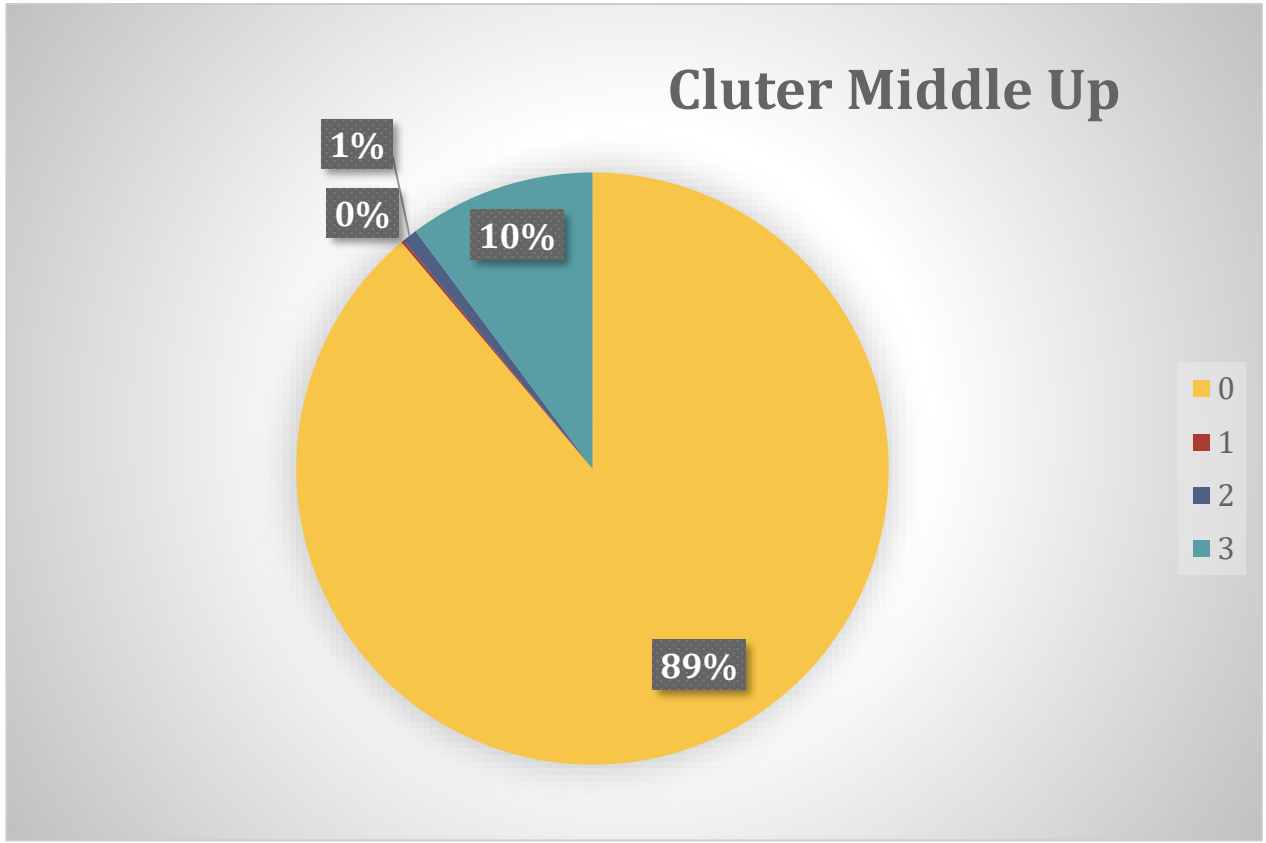
Cluster	Total Customer
0	1028
1	15
2	4474
3	14
4	136

Customer Segmentation Based on Middle-Up Class



In this segmentation, I used Jumlah and Harga_Satuan column to figuring out Purchasing Power of Customer. The summary is there are a lot customer who buy the things from range (0 – 6) for Jumlah, and 0 to 4 for Harga_Satuan

Percentage of Customer Segmentation Based on Middle Up Customer



Cluster	Total Customer
0	1840
1	3
2	17
3	211

Github Link