This is my adaptation of an article written by Baris Karaman on Towards Data Science.

I used his article as a way to get familiar with grouping customers based on their recency, frequency, and revenue—an incredibly powerful way to understand your customers and how they fit within specific segments.

This understanding can be used to guide future decisions and give non-technical stakeholders a quick, reliable, and easy-to-digest view of their customer base.

Baris's article can be found here:

https://towardsdatascience.com/data-driven-growth-with-python-part-2-customer-segmentation-5c019d150444

## RFM Segmentation

In this notebook, I'll be working through an orders dataset from a UK based online retailer and creating different customer segments based on their R (recency), F (frequency) and M (monetary value).

Recency can be thought of a few different ways, in todays notebook it will be shown as the number of days since a customers last purchase.

Frequency looks at how often a customer makes purchases within a specific time period.

Monetary value is a customers toal orders multiplied by their average order value, again within a specified time window. High revenue customers, usually goes hand in hand with recency and frequency.

Theoretically we'll end up with segements like...

## 1. Low Value

Customers who are not very frequent buyers and are less active. They genereate very low, or maybe even negative monetary value.

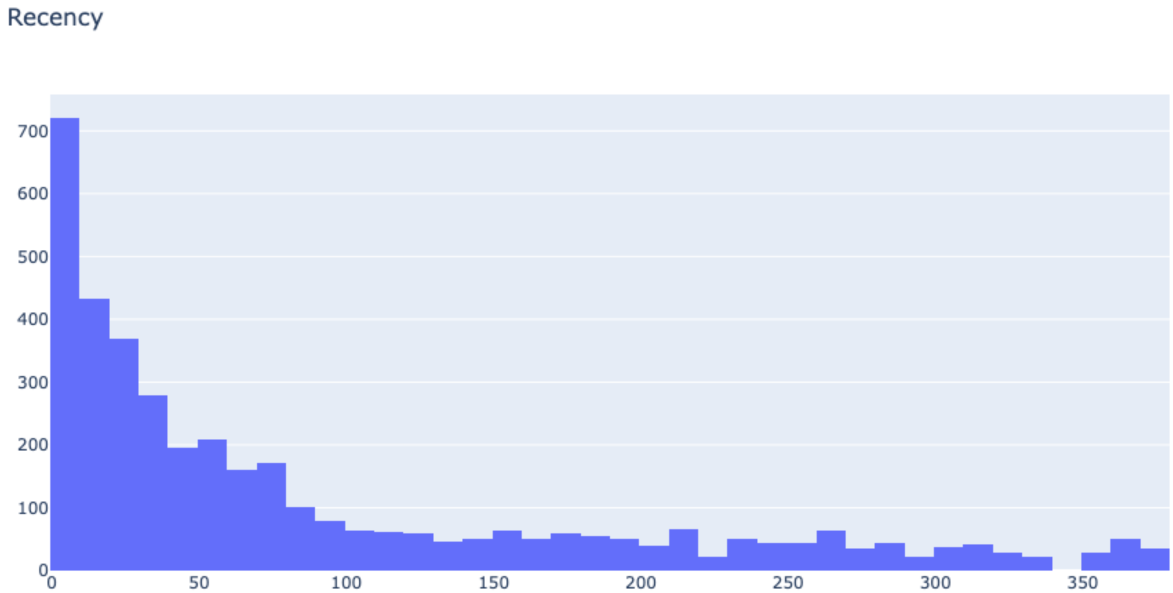
## 2. Medium Valie

Medium value customers tend to make purchases often and are fairly active. They generate solid revenue and are important to understand.

## 3. High Value

Frequent, high order volume customers. They generate a ton of revenue and are a group you never want to lose.

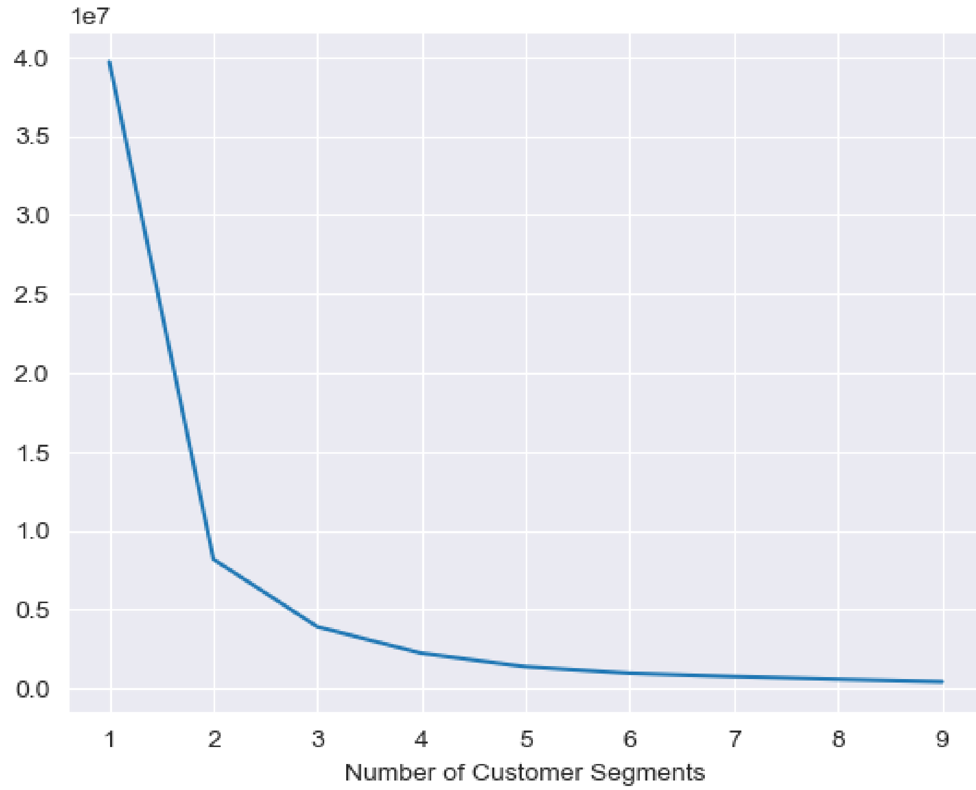
Let's look at recency first. We'll examine a customers most recent purchase find how long it's been since they've made that purchase. (measure in days)




| | CustomerID | Recency |
|---|---|---|
| **count** | 3950.000000 | 3950.000000 |
| **mean** | 15562.029367 | 90.778481 |
| **std** | 1576.848325 | 100.230349 |
| **min** | 12346.000000 | 0.000000 |
| **25%** | 14208.250000 | 16.000000 |
| **50%** | 15571.500000 | 49.000000 |
| **75%** | 16913.750000 | 142.000000 |
| **max** | 18287.000000 | 373.000000 |

So looking at our customers and their most recent purchase, the average recency (in days) is 90, and our median is 49 days. The longest it's been since a customer of ours had made a purchase is 373 days. We have our highest distrobution of recency within 0 - 9 days, meaning 720 customers have made a purchase within the last week and a half.

Now we're going to use machine learning, specifically K-means clustering, to group our customers into specific groups based upon their recency. Before we move forward, we need to determine how many clusters or segments we want to create. This is typically based up intuition and domain expertise, but since we don't have that right now, there are ways to ask our model what it thinks is best. Let's look.



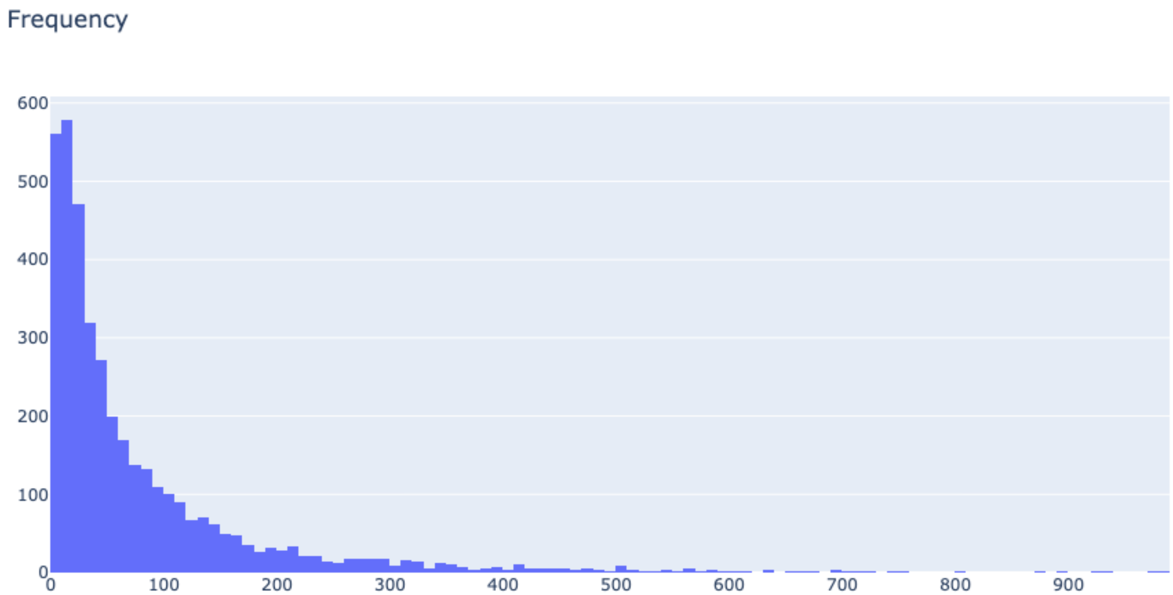

It looks like 3 is the optimal number of segments here. When looking at an elbow graph like this, we can make our decision based on the elbow in the graph and its corresponding x-value. A case could be made for two or three, and any real decision would be based on end business requirments. I'm actually going to choose four, as four customer segments makes the most sense to me.

| RecencyCluster | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **0** | 478.0 | 304.393305 | 41.183489 | 245.0 | 266.25 | 300.0 | 336.00 | 373.0 |
| **1** | 568.0 | 184.625000 | 31.753602 | 132.0 | 156.75 | 184.0 | 211.25 | 244.0 |
| **2** | 954.0 | 77.679245 | 22.850898 | 48.0 | 59.00 | 72.5 | 93.00 | 131.0 |
| **3** | 1950.0 | 17.488205 | 13.237058 | 0.0 | 6.00 | 16.0 | 28.00 | 47.0 |

We've ran our clustering model specifying four segments and assigned each customer to their appropriate segment. Segment 0 belongs to our least recent customers, and segment 3 belongs to our most recent.

We're going to do the same thing for frequency and monetary value.

Remember, recency saw how many days it's been since a customers last purchase. Frequency will look at how many unique purchases were made within a time period.

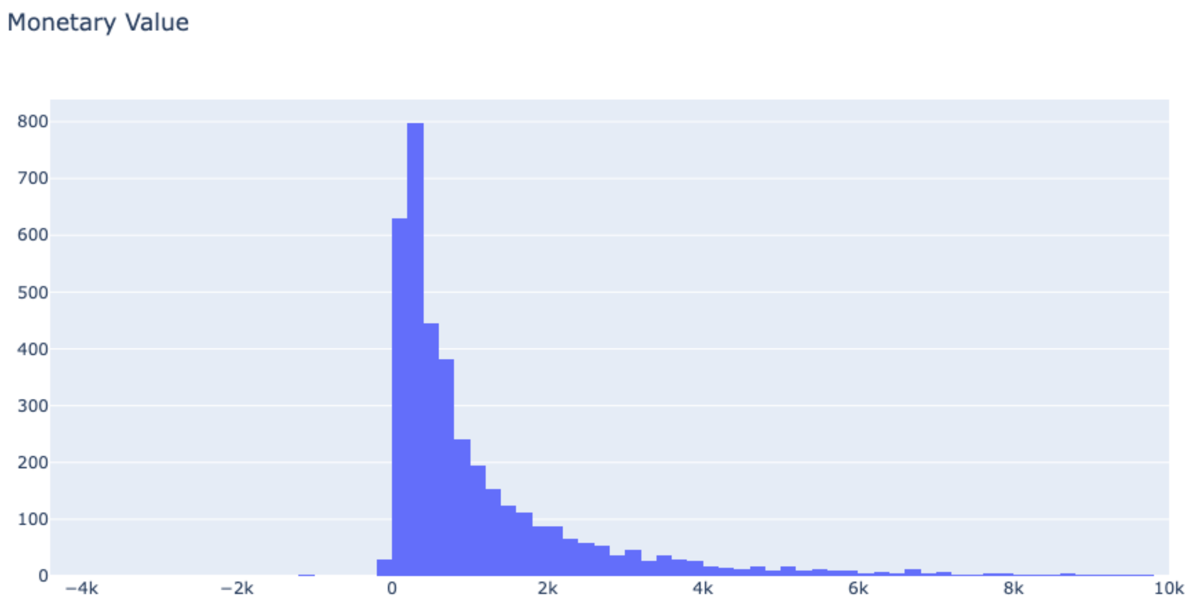



| FrequencyCluster | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 0 | 3496.0 | 49.525744 | 44.954212 | 1.0 | 15.0 | 33.0 | 73.0 | 190.0 |
| 1 | 429.0 | 331.221445 | 133.856510 | 191.0 | 228.0 | 287.0 | 399.0 | 803.0 |
| 2 | 22.0 | 1313.136364 | 505.934524 | 872.0 | 988.5 | 1140.0 | 1452.0 | 2782.0 |
| 3 | 3.0 | 5917.666667 | 1805.062418 | 4642.0 | 4885.0 | 5128.0 | 6555.5 | 7983.0 |

Segment 0 represents our least frequent customers, and segment three represents our most frequent customers.

You can see how many customers are in each segment, and what their average number of purchases is. For segment 3, the average number of purchases is 5917 an there are only 3 customers in this segment.

Now we'll do the same thing for monetary value. We can call if revenue from now on.




| RevenueCluster | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 0 | 3696.0 | 915.727558 | 936.627600 | -4287.63 | 264.2675 | 574.550 | 1262.0225 | 4435.79 |
| 1 | 226.0 | 7974.646062 | 3825.716457 | 4455.73 | 5308.3525 | 6662.095 | 9412.5725 | 25748.35 |
| 2 | 26.0 | 43736.679615 | 15866.891384 | 26626.80 | 29629.0950 | 38346.100 | 53859.2650 | 88125.38 |
| 3 | 2.0 | 221960.330000 | 48759.481478 | 187482.17 | 204721.2500 | 221960.330 | 239199.4100 | 256438.49 |

Same logic as before.

We now have our customers in three distinct segments based on their recency, frequency, and monetary value. Now we're going to combine the three to create an overall score, i.e., how well does a customer fit within all three segments. Could a customer have high revenue but low frequency and recency? We'll be able to answer questions like this after we find our combined score.

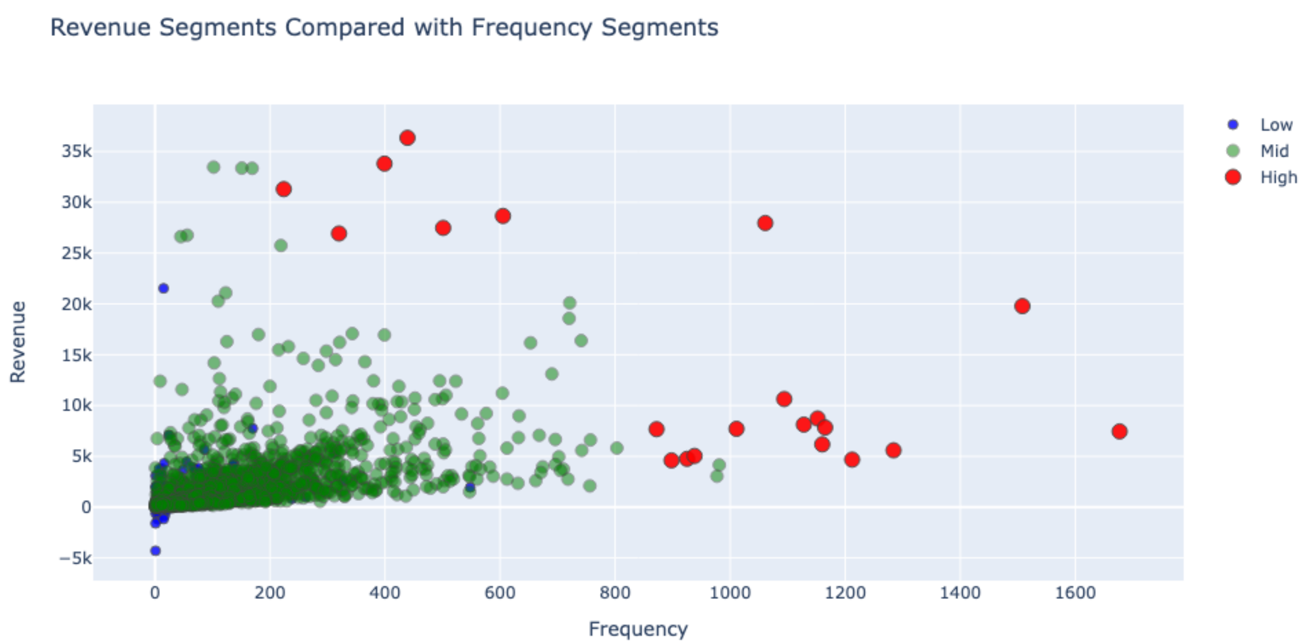
| OverallScore | Recency | Frequency | Revenue |
|---|---|---|---|
| 0 | 304.584388 | 21.995781 | 303.339705 |
| 1 | 185.362989 | 32.596085 | 498.087546 |
| 2 | 78.972856 | 47.060803 | 871.842586 |
| 3 | 20.684941 | 68.787318 | 1097.999703 |
| 4 | 14.627517 | 272.835570 | 3607.121678 |
| 5 | 9.600000 | 374.103448 | 9382.324276 |
| 6 | 8.038462 | 901.307692 | 22663.667308 |
| 7 | 1.857143 | 1272.714286 | 103954.025714 |
| 8 | 1.333333 | 5917.666667 | 42177.930000 |

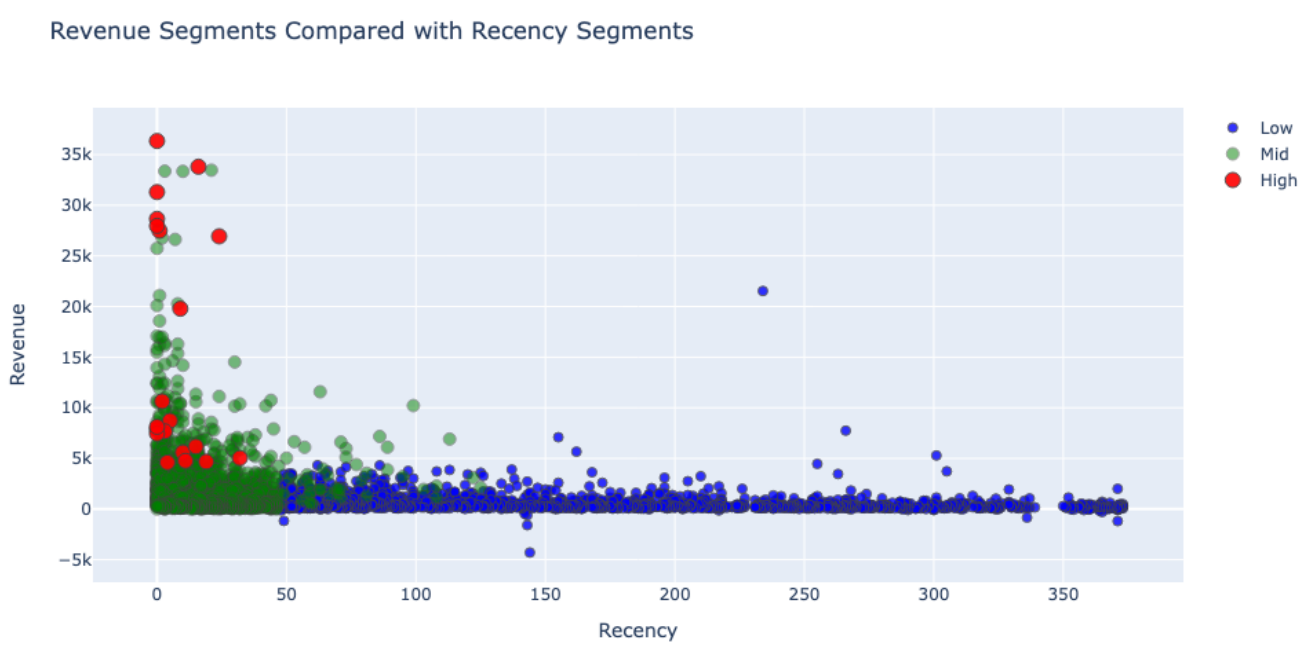
Remember that our highest segment was three for each group, so the maximum possible score is nine if they scored three across each cluster.

Eight seems to be the highest score here, with customers who scored eight being our most valuable customers. Curiously, though, the average revenue seems to be way higher for cluster seven. Perhaps a customer who spent a massive amount a while ago is causing that.

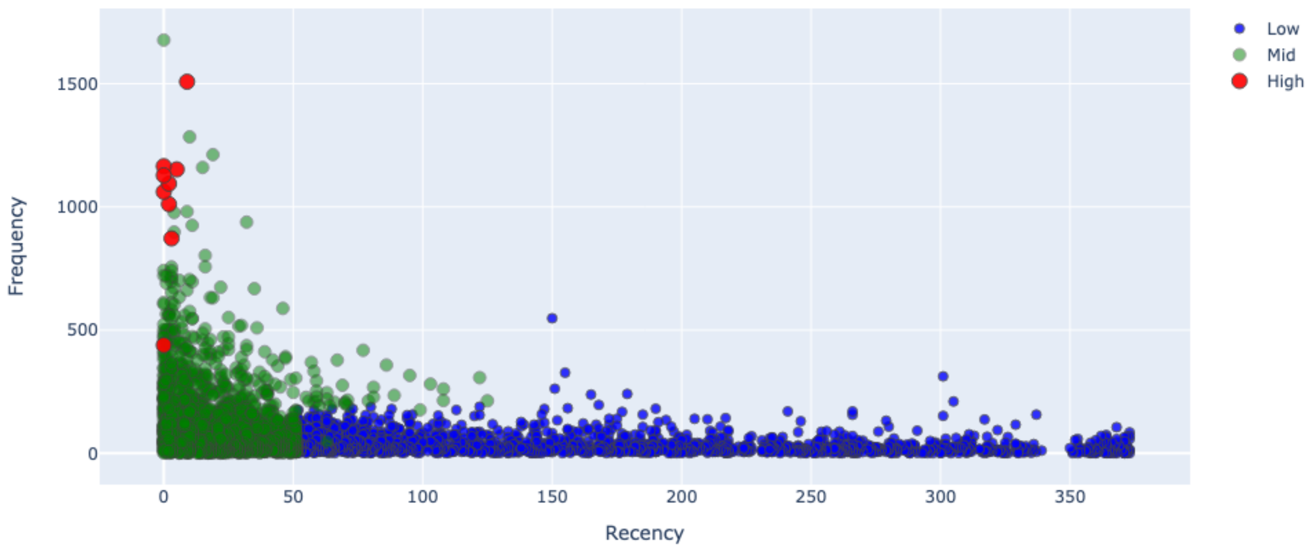
Let's make things easier to understand by throwing our customers into three segments:

1. Low value: corresponding to a score of 0 - 2.
2. Medium value: corresponding to a score of 3 - 5.
3. High value: corresponding to a score of 6+



Frequency Segments Compared with Recency Segments




The takeaways here are fairly intuitive:

1. You can see high-value customers tend to make more purchases and drive more revenue. (graph 1)
2. You can see high-value customers tend to be the most recent buyers and have higher revenue. (graph 2)
3. You can see high-value customers tend to be the most recent and most frequent. (graph 3)

Using these segments, we can begin to take action. We can create stategies to optimize our high value customers, and strategies to improve RFM among mid and low value customers.