

DATA analysis on CRM Dataset

**Booster Box company
Farzad Imanpour Sardroudi**

CONTENTS

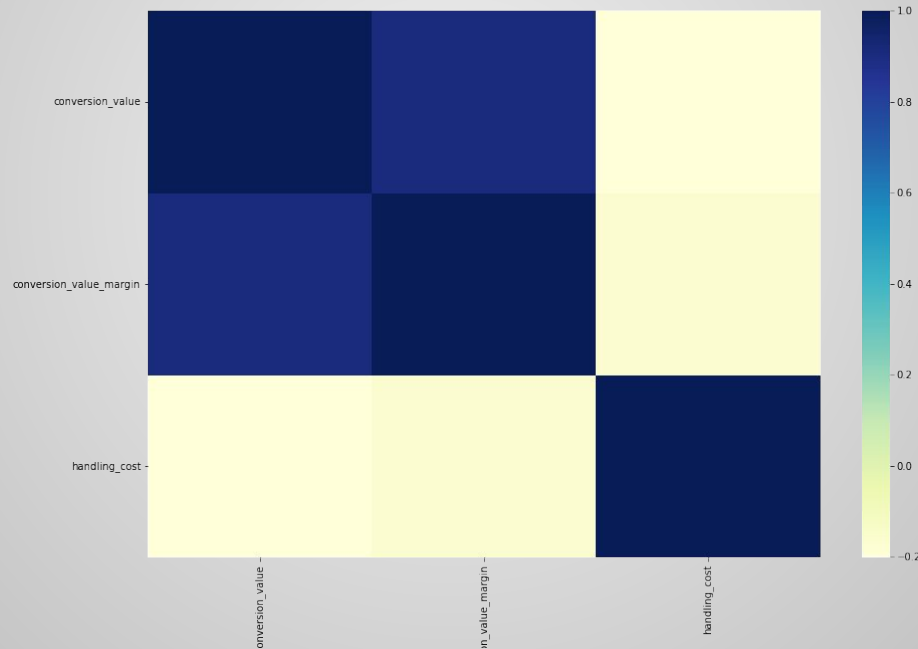


Pre-processing:

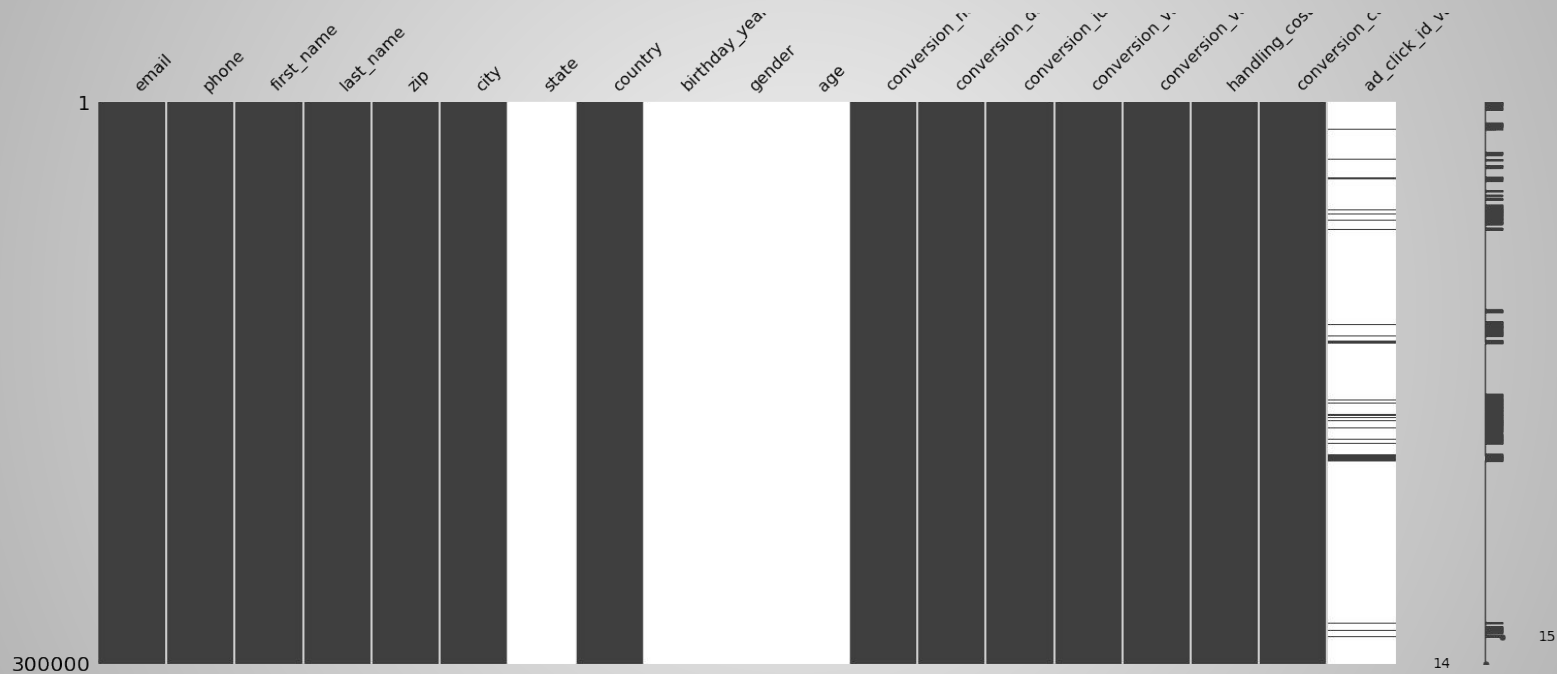
- Data quality assessment
 - Mismatched data types
 - Mixed data values
 - Data outliers
 - Missing data
- Data cleaning
 - Ignore the tuples
 - Manually fill in missing data
- Data transformation
 - Aggregation
 - Normalization
 - Feature selection
 - Discretization

HeatMap of Correlations

used to find the pairwise correlation of all columns in the dataframe. Any Na values are automatically excluded. For any non-numeric data type columns in the dataframe it is ignored. Strengthen your foundations with the Python Programming Foundation Course and learn the basics



Missing values

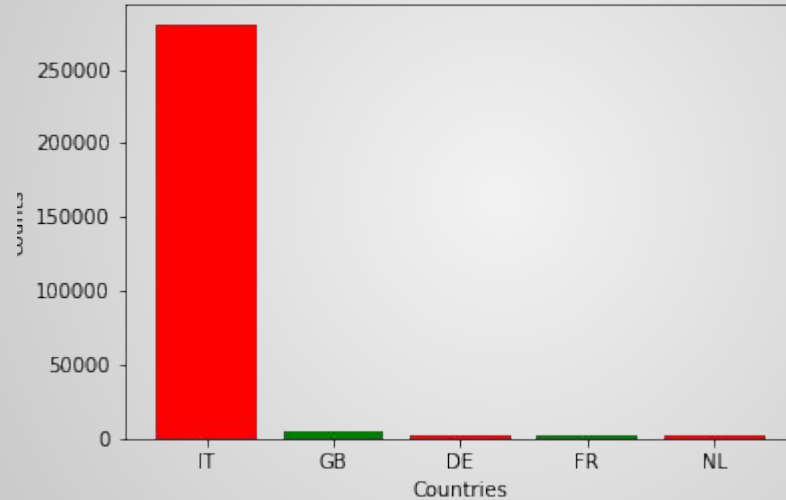


Merging columns and optimize dataframe

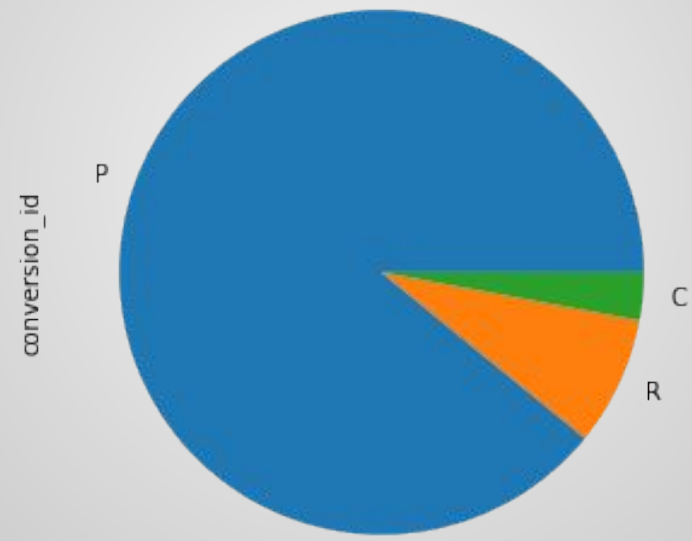
- Merging email, phone, first_name and last_name to generate a unique ID for each customer
- Merging conversion_name and conversion_id
 - For conversion_name we have 2 options:
 - 1-just encoding the 3 types we have(We had 4 types but 2 of them were same, and I merged them)
 - 2-adding the first letter of it to ID, we can remove this column, and it would be optimized

	customer_id	conversion_id
0	62680	P918183
1	124148	C917656
2	72157	R905587
3	136027	P918391
4	158402	C917776

The distribution of Customers according to countries



numbers of total purchased, returned and Canceled

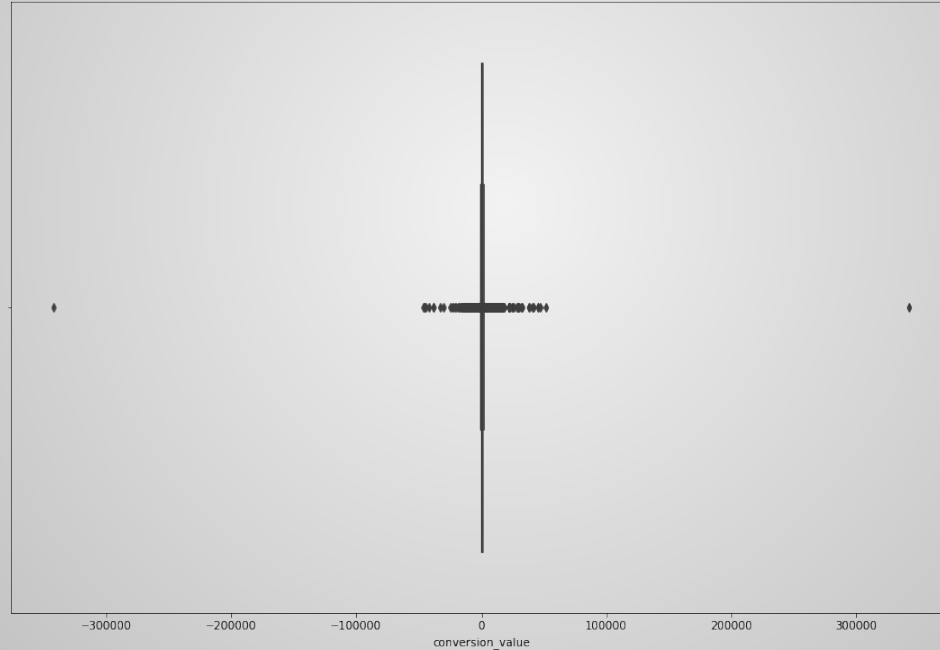


Outlier detection

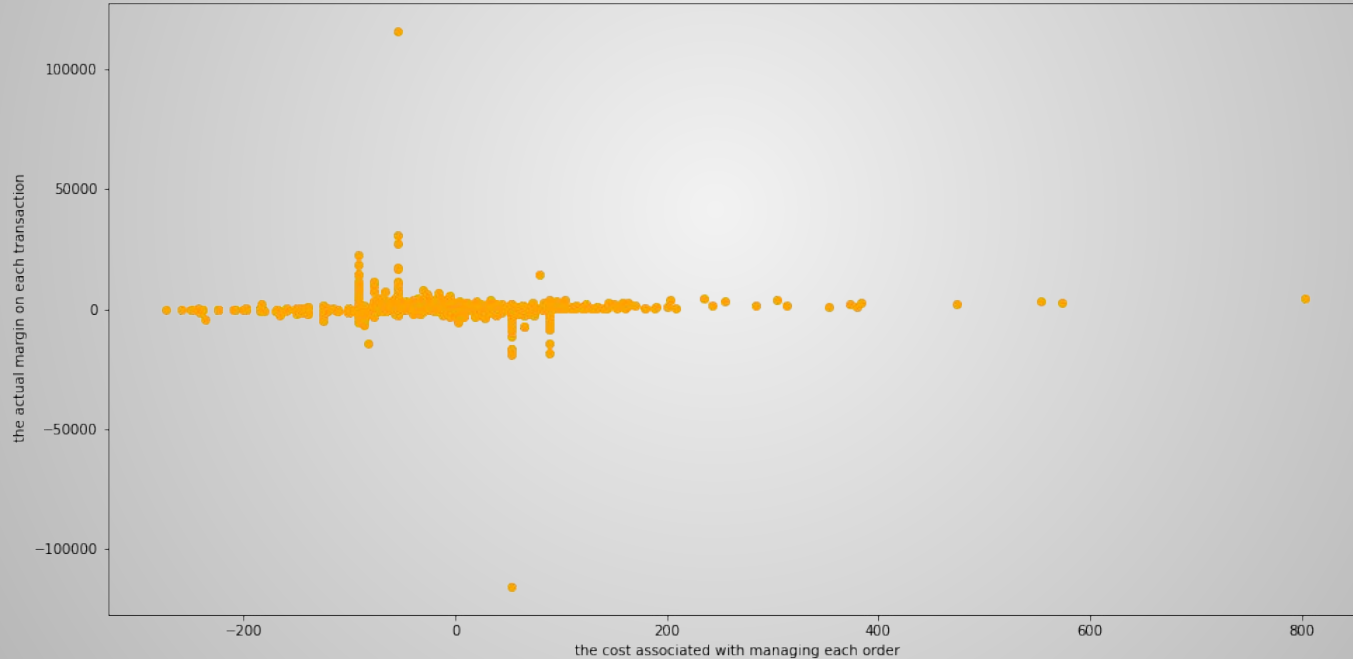
The interquartile range (IQR), also called the midspread or middle 50%, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, $IQR = Q3 - Q1$.

It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outliers.

Below plot shows points which are so far, these are outliers as there are not included in the box of other observation i.e. no where near the quartiles.



Looking at the plot below, we can see most of the data points are lying center side, but there are points which are far from the population (conversion_value, conversion_value_margin).



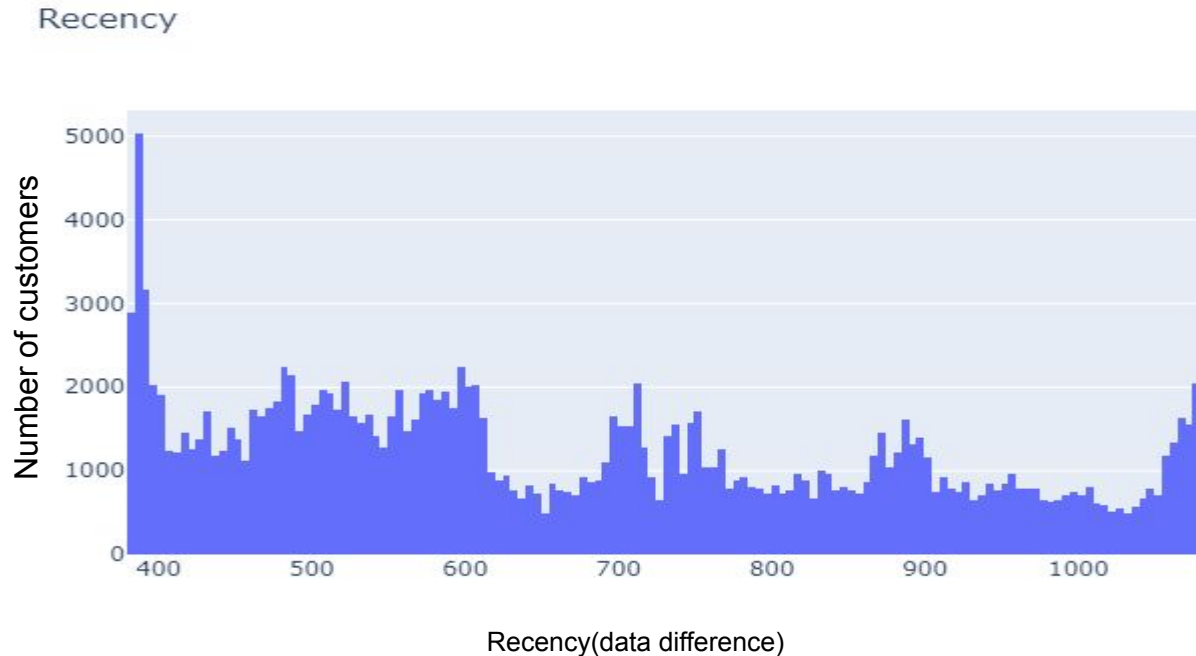
RFM analysis (Recency - Frequency - Monetary)

- Can not treat every customer the same way
- Customers who use your platform have different needs, and they have their own different profile.
 - You should adapt your actions depending on that
- For increasing retention rate, a good way is segmentation
- ★ we are going to implement one of them to our business which is RFM.

Recency

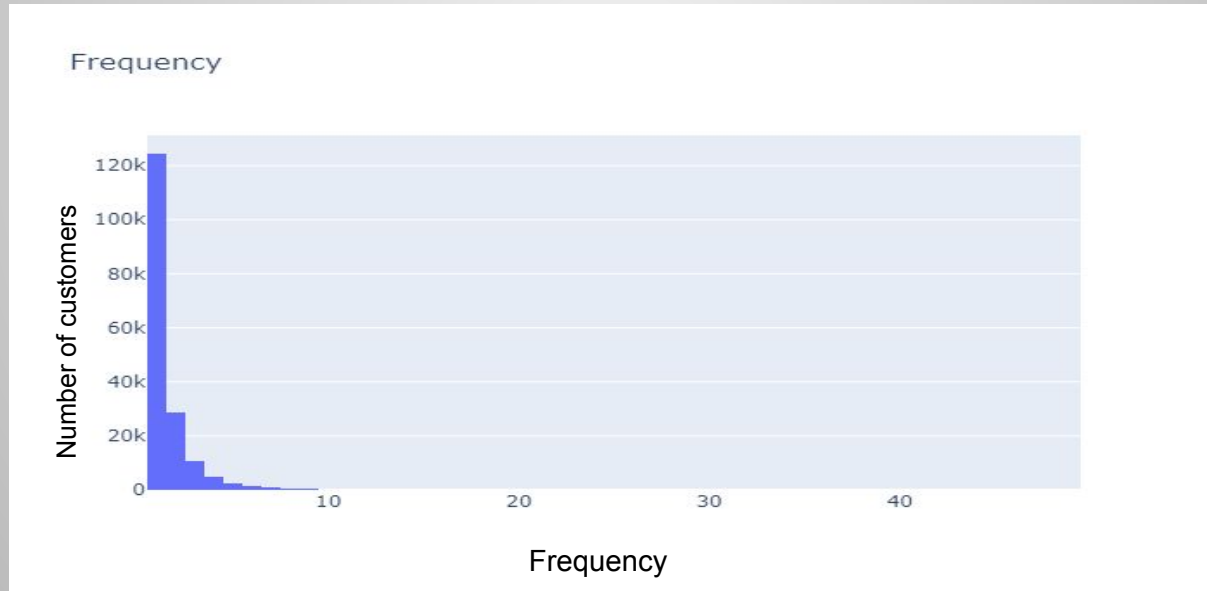
To calculate recency, we need to find out the most recent purchase date of each customer and see how many days they are inactive for. We will apply K-means clustering to assign customers a recency score.

Our code snippet above has a histogram output to show us how is the distribution of recency across our customers.



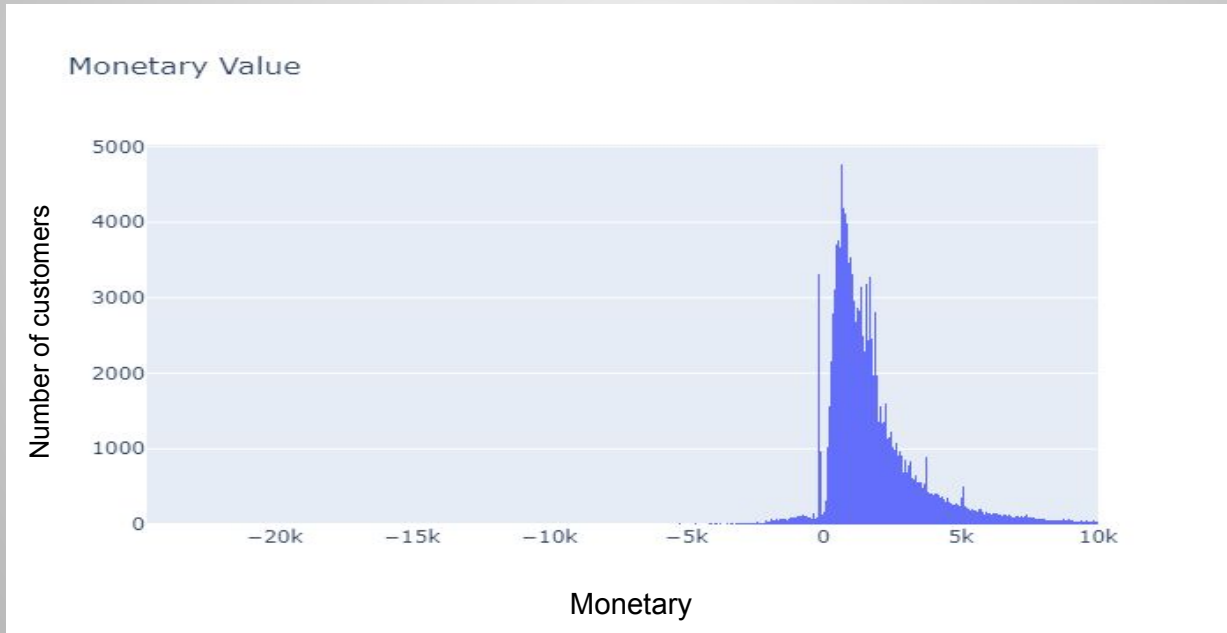
Frequency

We need to find the total number of orders for each customer. First calculate this and see how frequency look like in our customer database



Monetary

How our customer database looks like when we cluster them based on revenue. We will calculate revenue for each customer



Create RFM Score

Discretizing variables into equal-sized buckets based on rank or based on sample quantiles.

	Recency	Frequency	Monetary	RecencyScore	FrequencyScore	MonetaryScore	RFM_SCORE	Segment
customer_id								
53700	393	4	8907.550000	5	5	5	555	High-Value
110042	384	3	5652.200000	5	5	5	555	High-Value
5239	404	4	9041.440000	5	5	5	555	High-Value
83775	417	4	4364.330000	5	5	5	555	High-Value
16701	389	3	6788.830000	5	5	5	555	High-Value

To keep things simple, better we name these scores

Theoretically we will have segments like below:

- Low Value: Customers who are less active than others, not very frequent buyer/visitor and generates very low - zero - maybe negative revenue.
- Mid-Value: In the middle of everything. Often using our platform (but not as much as our High Values), fairly frequent and generates moderate revenue.
- High Value: The group we don't want to lose. High Revenue, Frequency and low Inactivity.

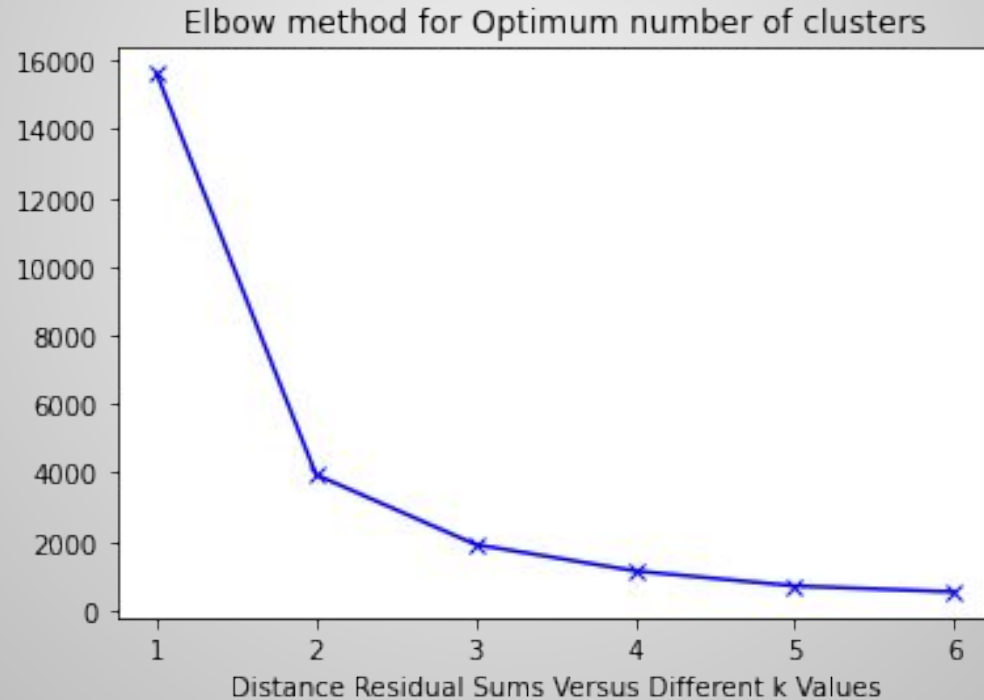
K-Means

We are going to apply K-means clustering to assign a recency score. But we should tell how many clusters we need to K-means algorithm. To find it out, we will apply Elbow Method. Elbow Method simply tells the optimal cluster number for optimal inertia.

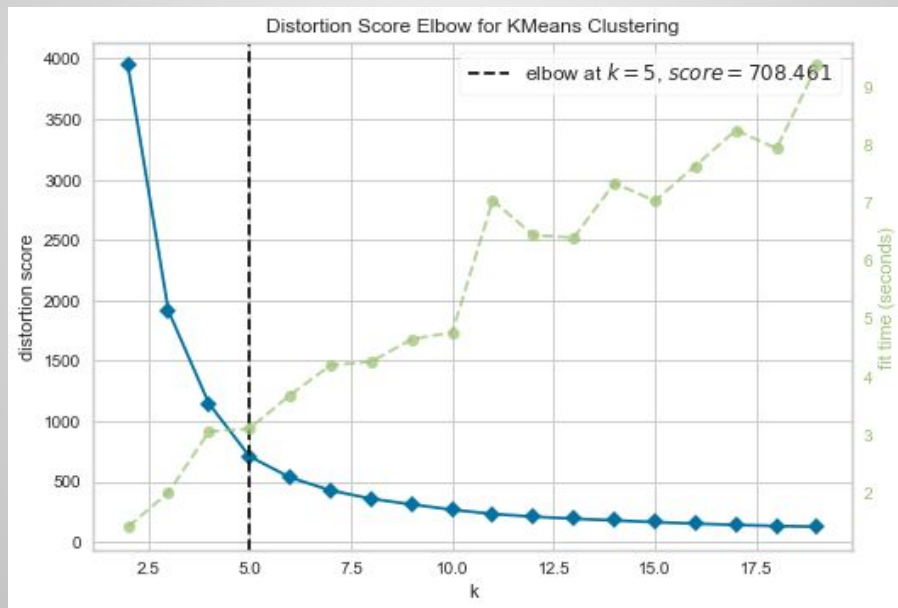
As the methodology, we need to calculate Recency, Frequency and Monetary Value and apply unsupervised machine learning to identify different groups (clusters) for each.

We should tell how many clusters we need to K-means algorithm. To find it out, we will apply Elbow Method. Elbow Method simply tells the optimal cluster number for optimal inertia.

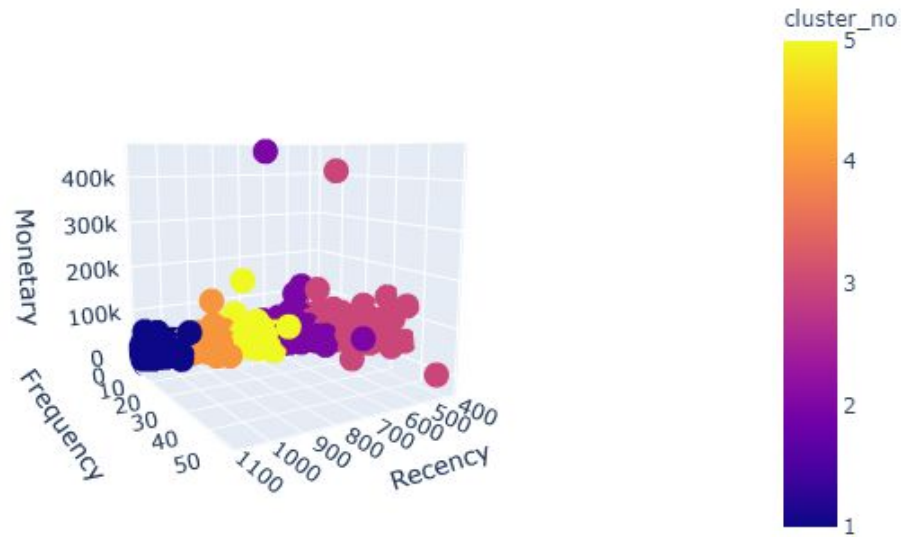
A good model is one with low inertia AND a low number of clusters .



The KElbowVisualizer implements the “elbow” method to help data scientists select the optimal number of clusters by fitting the model with a range of values for K. If the line chart resembles an arm, then the “elbow” (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. In the visualizer, “elbow” will be annotated with a dashed line.



Plotting the clusters



Customer Lifetime Value

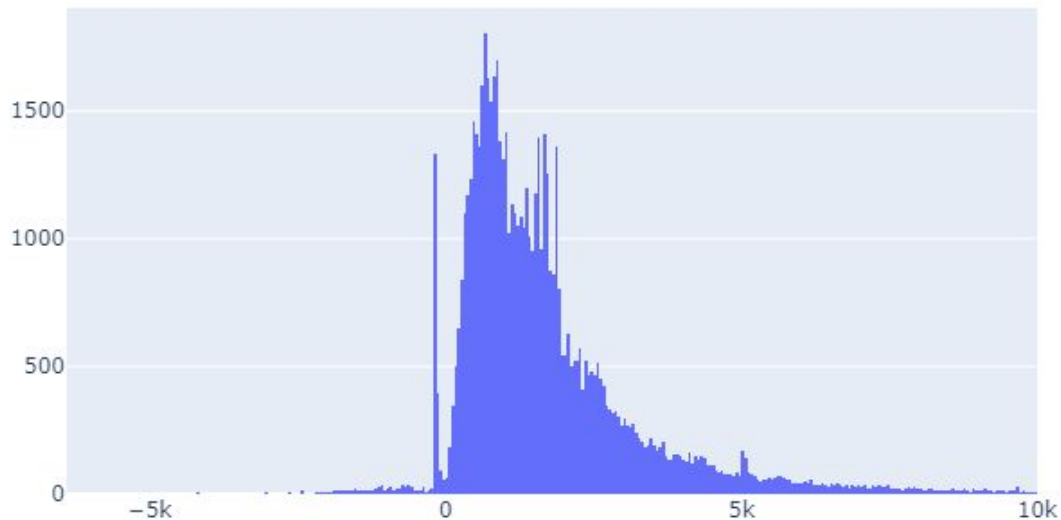
We invest in customers (acquisition costs, offline ads, promotions, discounts & etc.) to generate revenue and be profitable. Naturally, these actions make some customers super valuable in terms of lifetime value, but there are always some customers who pull down the profitability. We need to identify these behavior patterns, segment customers, and act accordingly.

Calculating Lifetime Value is the easy part. First, we need to select a time window. It can be anything like 3, 6, 12, 24 months. By the equation below, we can have Lifetime Value for each customer in that specific time window.

This gives us the historical lifetime value. If we see some customers having very high negative lifetime value historically, it could be too late to take an action. At this point, we need to predict the future with machine learning:

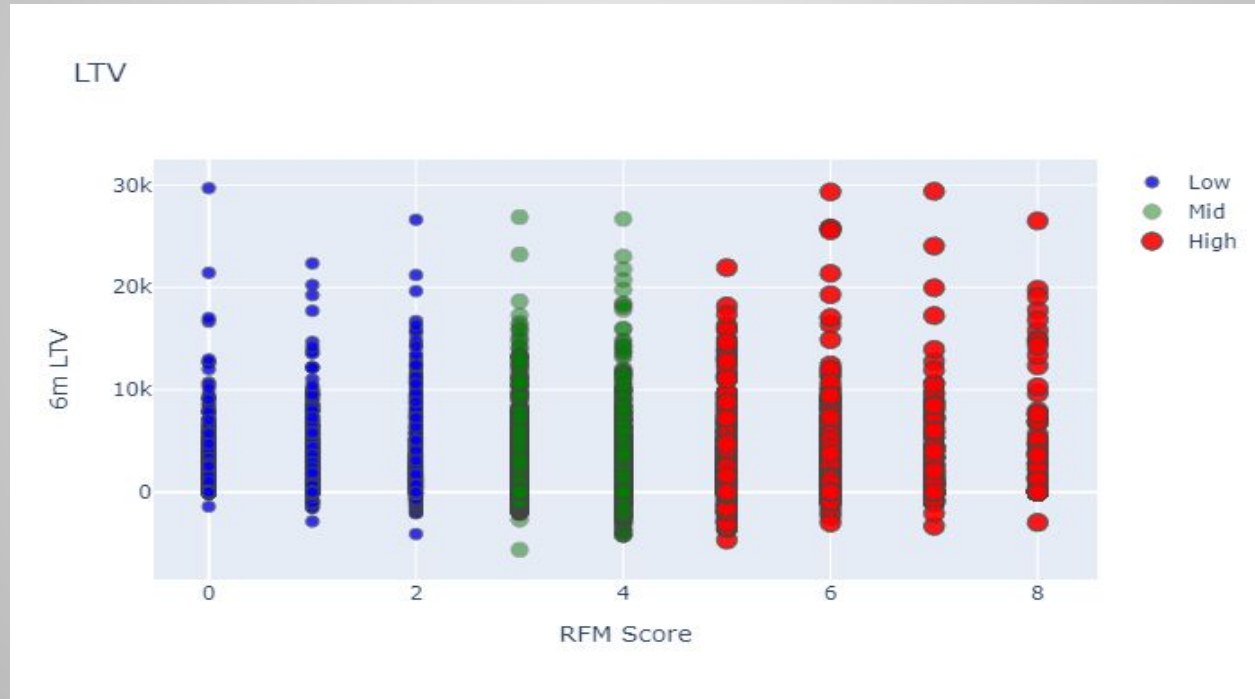
Calculate 6 months LTV for each customer

6m Revenue



Merge our 3 months and 6 months dataframes to see correlations between LTV and the feature set we have.

Positive correlation is quite visible here. High RFM score means high LTV.



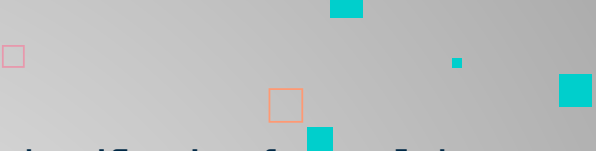
LTV itself is a regression problem. A machine learning model can predict the \$ value of the LTV. But here, we want LTV segments. Because it makes it more actionable and easy to communicate with other people. By applying K-means clustering, we can identify our existing LTV groups and build segments on top of it.

	count	mean	std	min	25%	50%	75%	max
LTVCluster								
0	27282.00	3.46	237.93	-5639.73	0.00	0.00	0.00	841.32
1	3217.00	1681.96	570.18	841.90	1190.55	1605.87	2116.00	2877.33
2	1291.00	4080.28	872.66	2879.03	3331.48	3921.24	4727.02	6062.06

2 is the best with average 4k LTV whereas 0 is the worst with 3.

Then we do some processes before creating model:

- convert categorical columns to numerical columns.
- split our feature set and label (LTV) as X and y. We use X to predict y.
- create Training and Test dataset. Training set will be used for building the machine learning model.




We used a quite strong ML library called XGBoost to do the classification for us. It has become a multi classification model since we had 3 groups (clusters). Let's look at the initial results:

We got

- Accuracy of XGB classifier on training set: 0.86
- Accuracy of XGB classifier on test set: 0.86

Precision and recall are acceptable for all, but we should work more on both 2nd and 3rd cluster by trying other models or adding new features.

Now we have a machine learning model which predicts the future LTV segments of our customers. We can easily adapt our actions based on that.



Other steps that we can work on it are:

1. Churn Prediction
2. Predicting Next Purchase Day
3. Predicting Sales
4. Market Response Models
5. Uplift Modeling
6. A/B Testing Design and Execution

For other ways of segmentation, we can:

- **Demographic segmentation**: where an organization's target market is segmented based on demographic variables : age, gender, income.
- **Psychographic segmentation**: breaks down your customer groups into segments that influence buying behaviors: beliefs, values, lifestyle, social status.
- **Behavioral segmentation**: sorting customers based on the behaviors they exhibit. These behaviors include the types of products and content they consume, and the cadence of their interactions with an app, website, or business
- **geographic segmentation**: marketing strategy used to target products or services at people who live in

Thank you.