Segmentation of e-commerce website customers

Adonija ZIO

Outline


Executive Summary

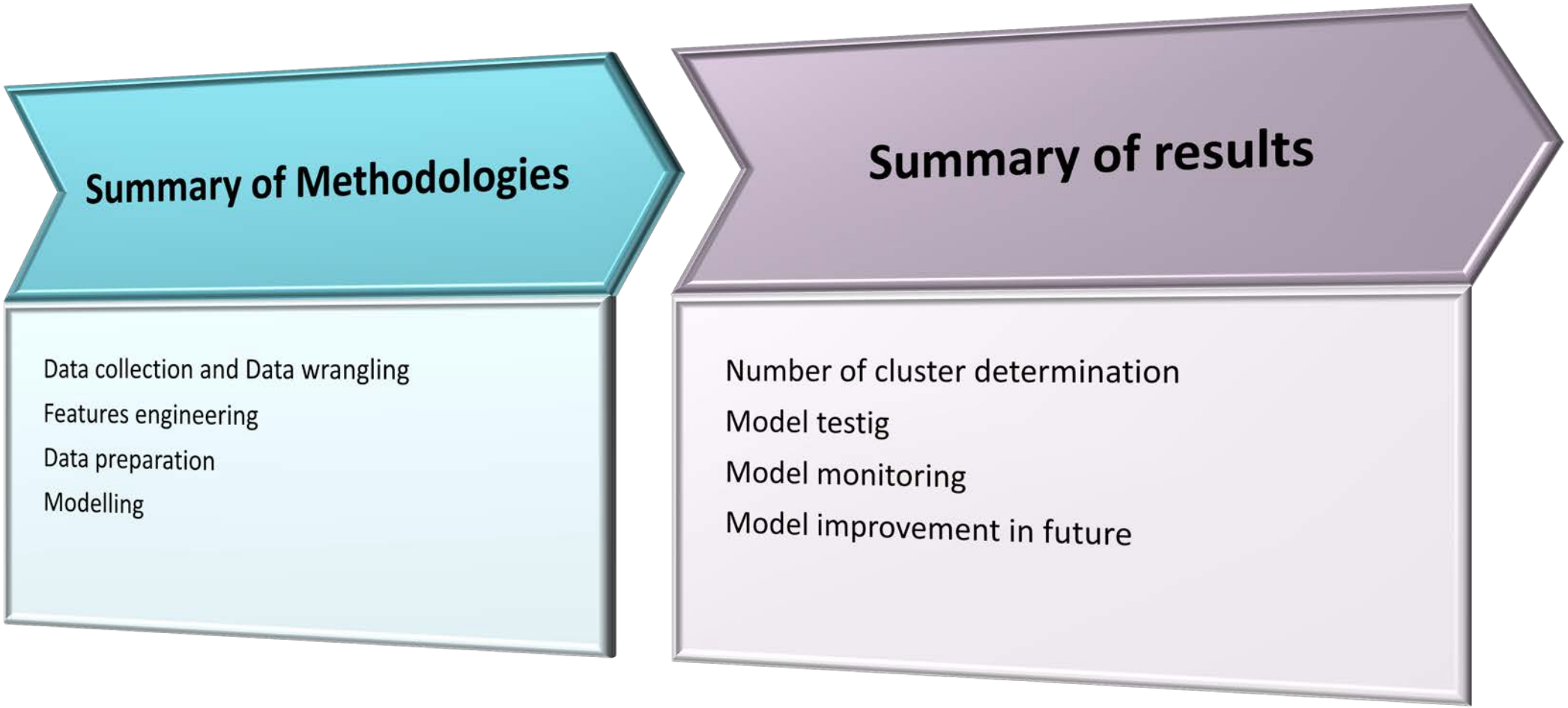

Introduction


Methodology

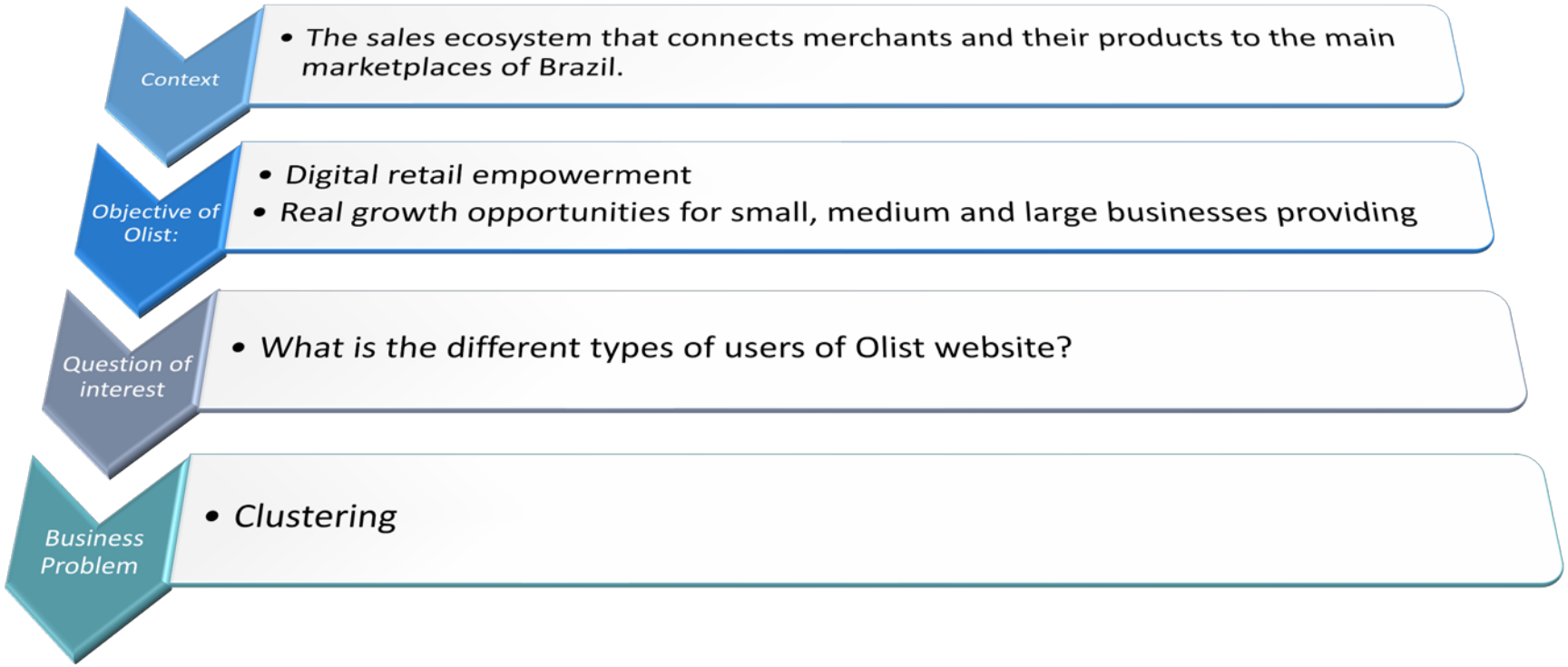

Results


Conclusion

Executive Summary



Introduction



Methodology

DATASETS PREVIEW

DATASETS MERGING

FEATURE ENGINEERING

EXPLORATORY DATA ANALYSIS

MODEL TESTING

MONITORING

Data understanding and data wrangling

Datasets preview

- Understanding of each datasets
- Relation between the different datasets
- Dealing with missing values

Datasets merging

- Droping duplicated values
- Merging datasets with merging keys
- Droping outliers

	datasets	columns	total_rows	total_cols	total_duplicate	total_null	null_cols
0	customers	customer_id, customer_unique_id, customer_zip_code_prefix, customer_city, customer_state	99441	5	0	0	
1	geolocation	geolocation_zip_code_prefix, geolocation_lat, geolocation_lng, geolocation_city, geolocation_state	738332	5	0	0	
2	items	order_id, order_item_id, product_id, seller_id, shipping_limit_date, price, freight_value	112650	7	0	0	
3	payments	order_id, payment_sequential, payment_type, payment_installments, payment_value	103886	5	0	0	
4	reviews	review_id, order_id, review_score, review_comment_title, review_comment_message, review_creation_date, review_answer_timestamp	99224	7	0	0	
5	orders	order_id, customer_id, order_status, order_purchase_timestamp, order_approved_at, order_delivered_carrier_date, order_delivered_customer_date, order_estimated_delivery_date	99441	8	0	4908	order_approved_at, order_delivered_carrier_date, order_delivered_customer_date
6	products	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	32951	9	0	2448	product_category_name, product_name_lenght, product_description_lenght, product_photos_qty, product_weight_g, product_length_cm, product_height_cm, product_width_cm
7	sellers	seller_id, seller_zip_code_prefix, seller_city, seller_state	3095	4	0	0	
8	category	product_category_name, product_category_name_english	71	2	0	0	

Features engineering

- Features creation
  - Distance calculation
  - Delevery delay
  - Ratio calculation

- Data aggregation
  - Customer profil construction
  - Feature reclaculation

Results

Exploratory Data Analysis

- Evolution of variables over time
- Numerical features visualization

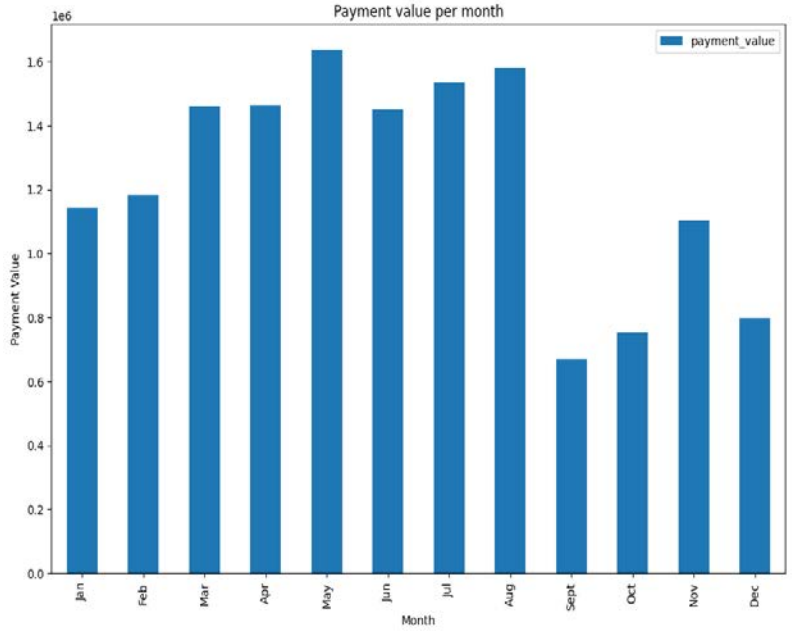
Model testing

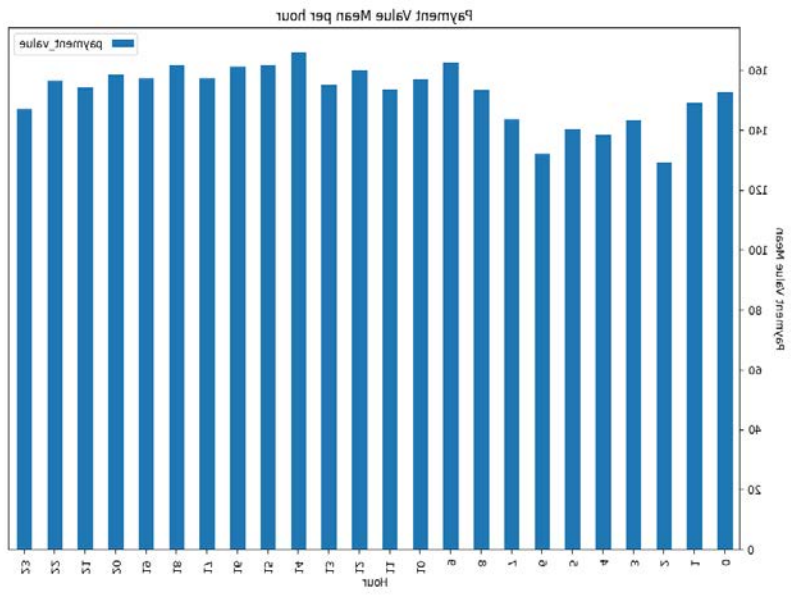
- Number of cluster selection
- Modelling

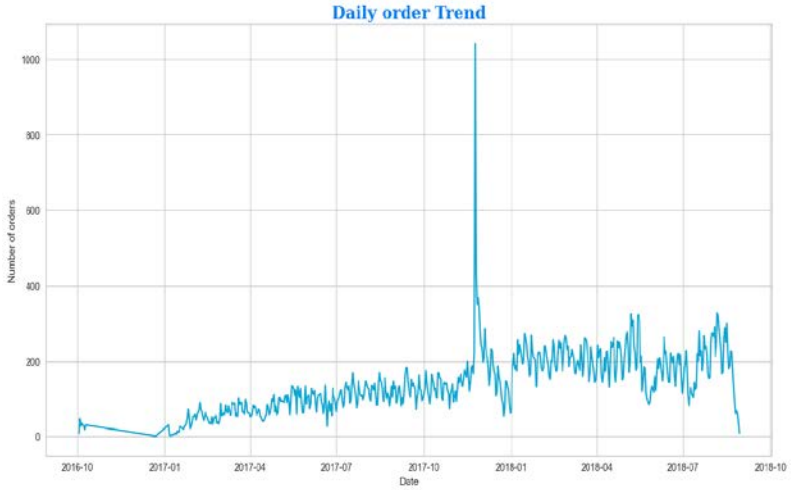
Model Monitoring

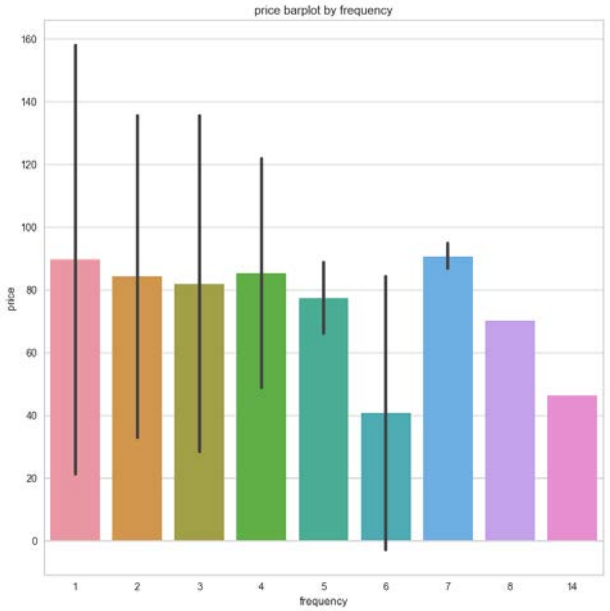
- Model simulation
- Update deadline

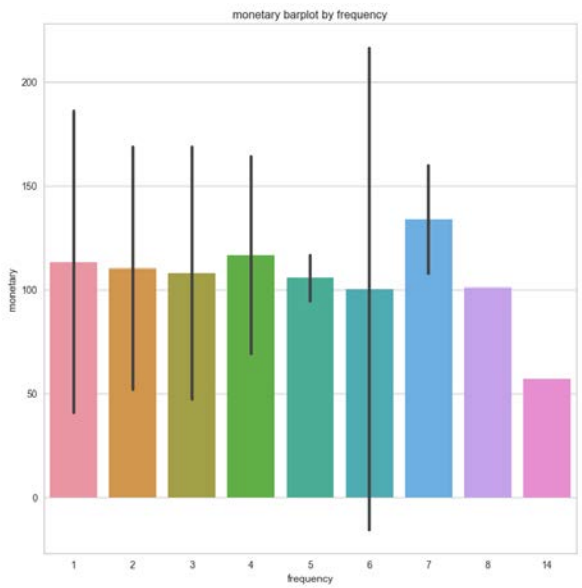
Exploratory analysis: Evolution of variables over time

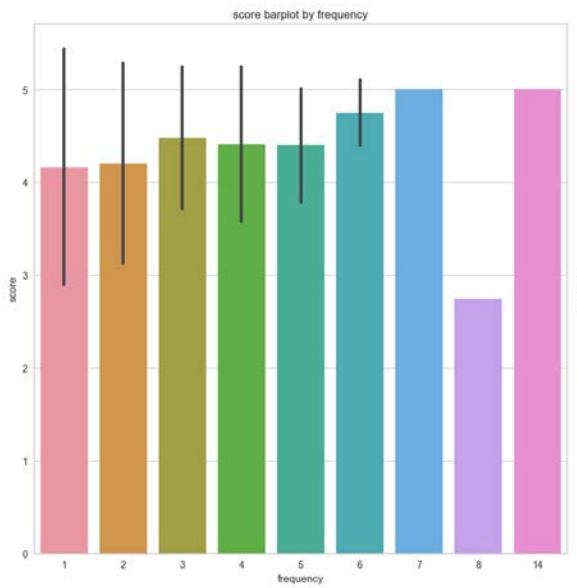


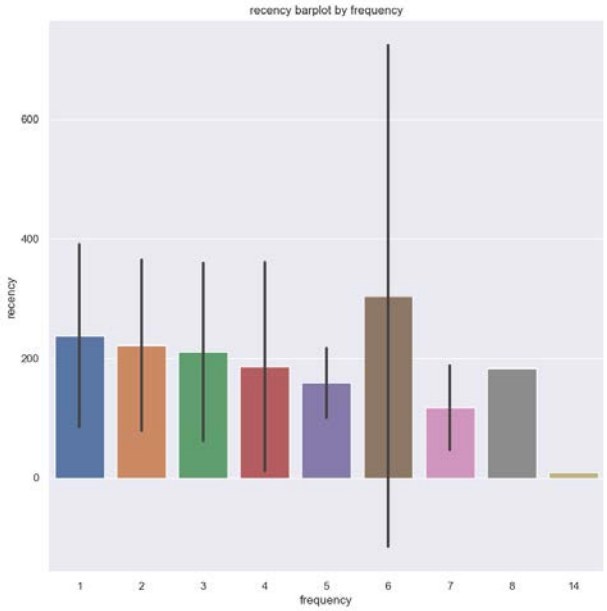






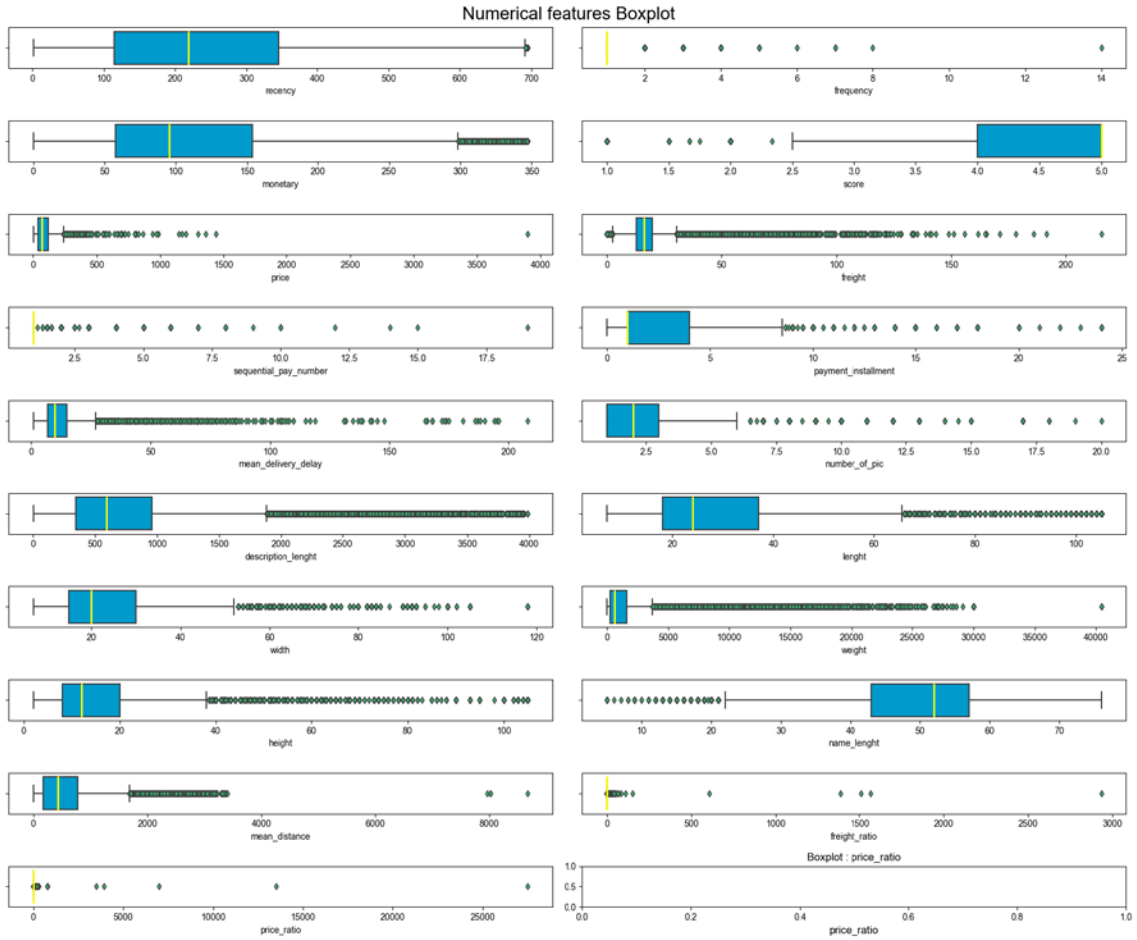


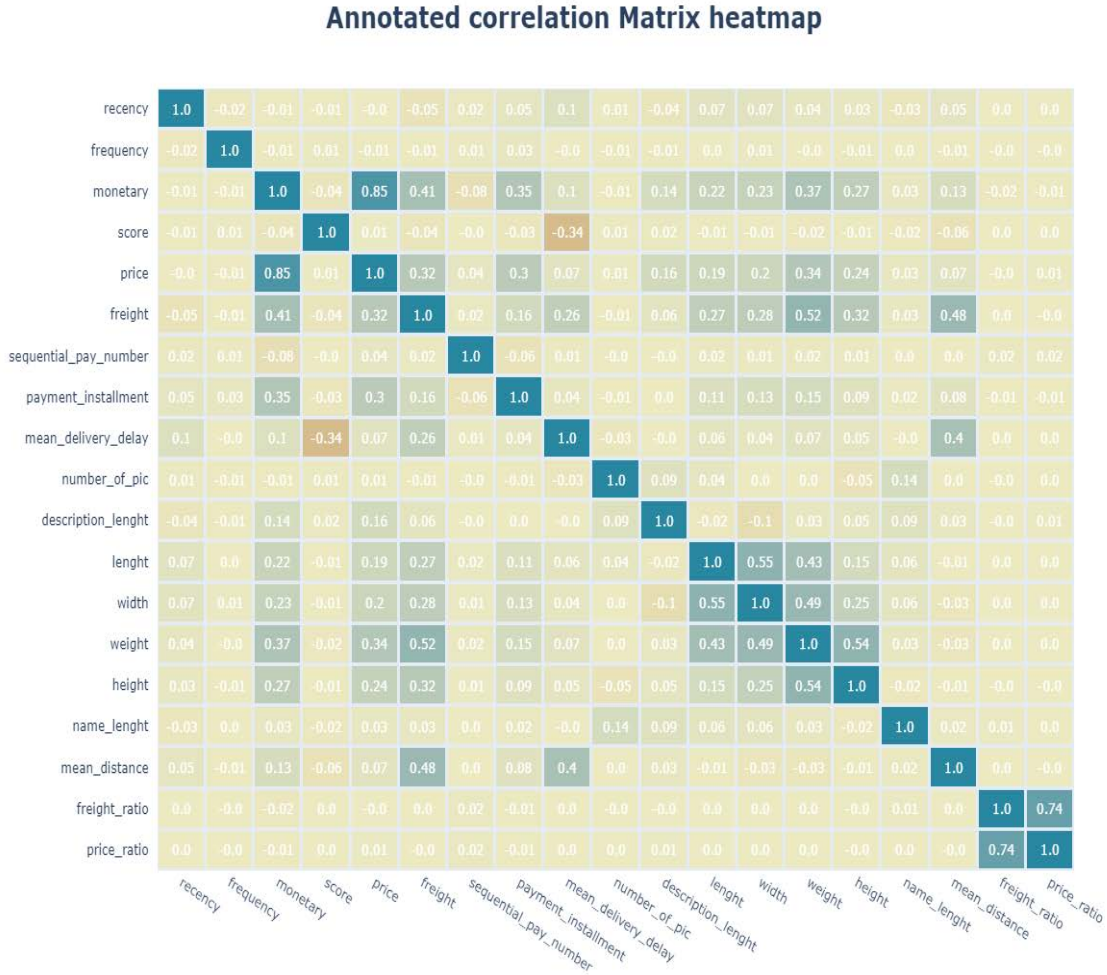




Exploratory analysis: Relationship between features aggregation by costumer

Exploratory analysis: Relationship between features aggregation by costumer





Modelling

Model building

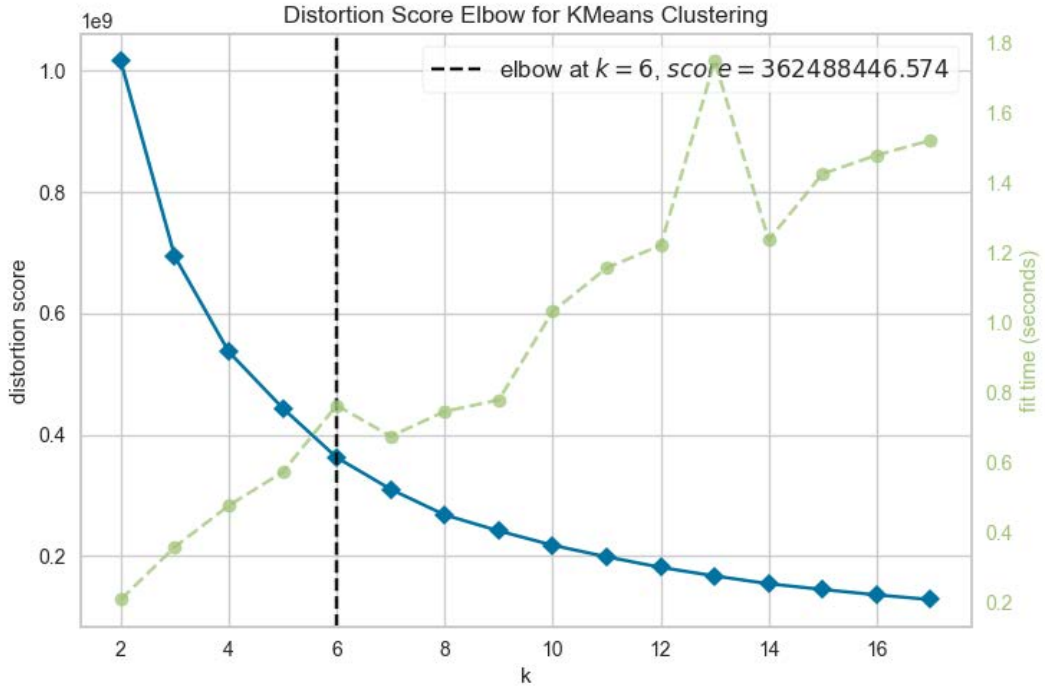
Initial model
Optimal K determination
Model evaluation

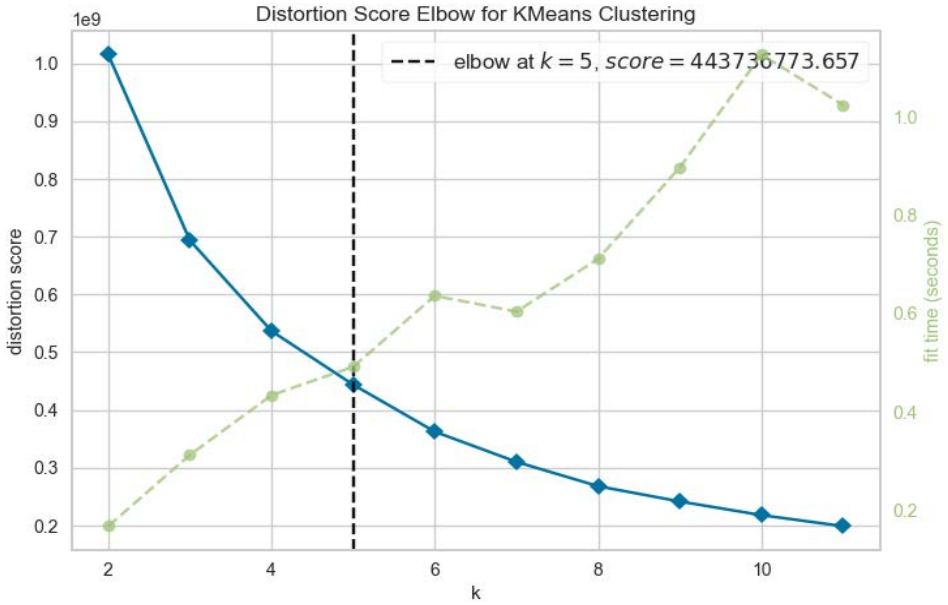
Model testing

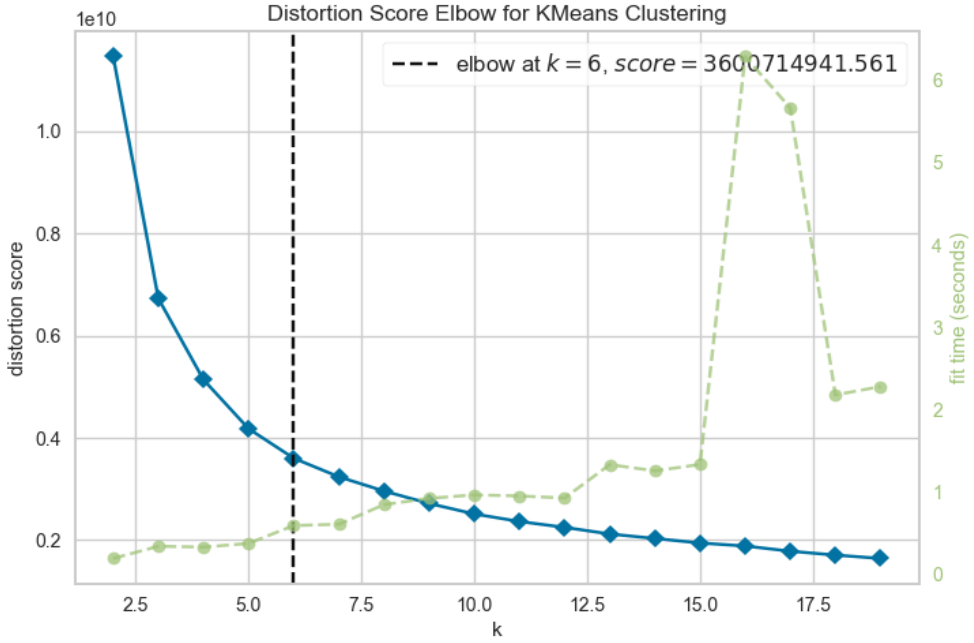
Model selection

Model monitoring

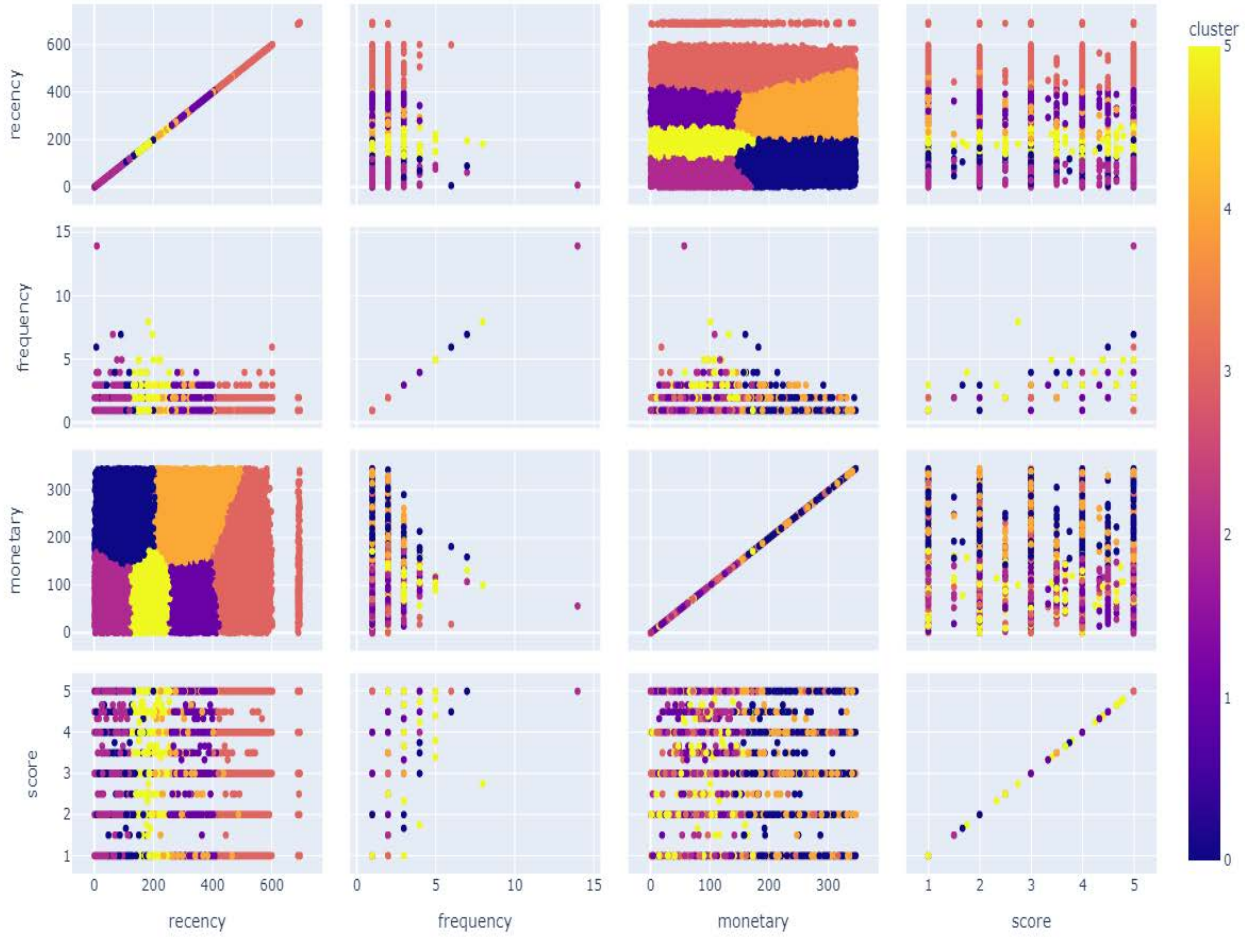
Optimal cluster

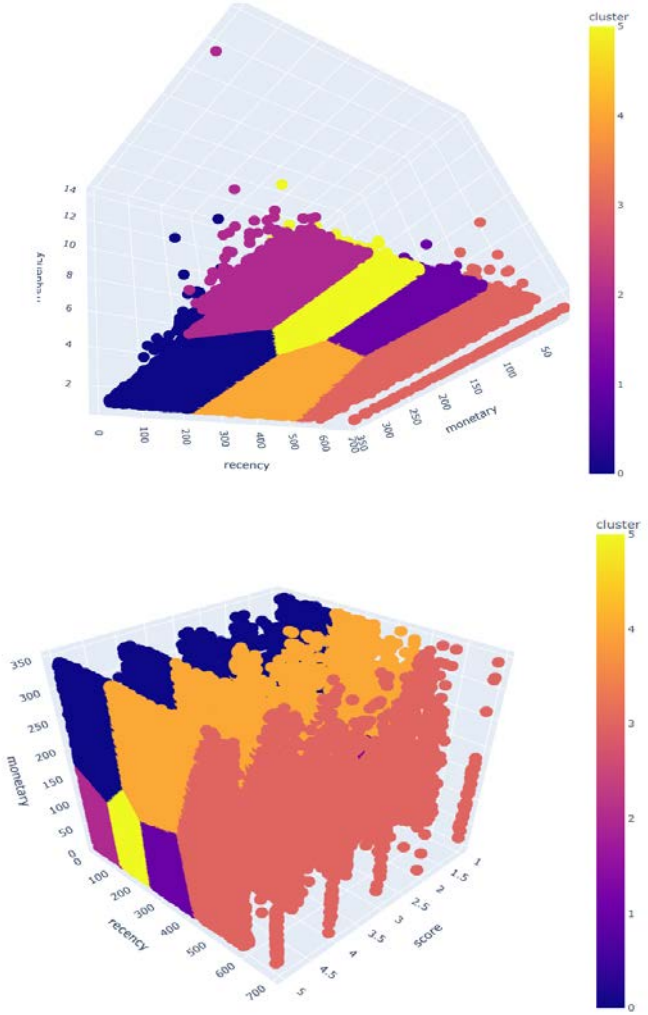






RFM and RFMS clustering with Kmeans



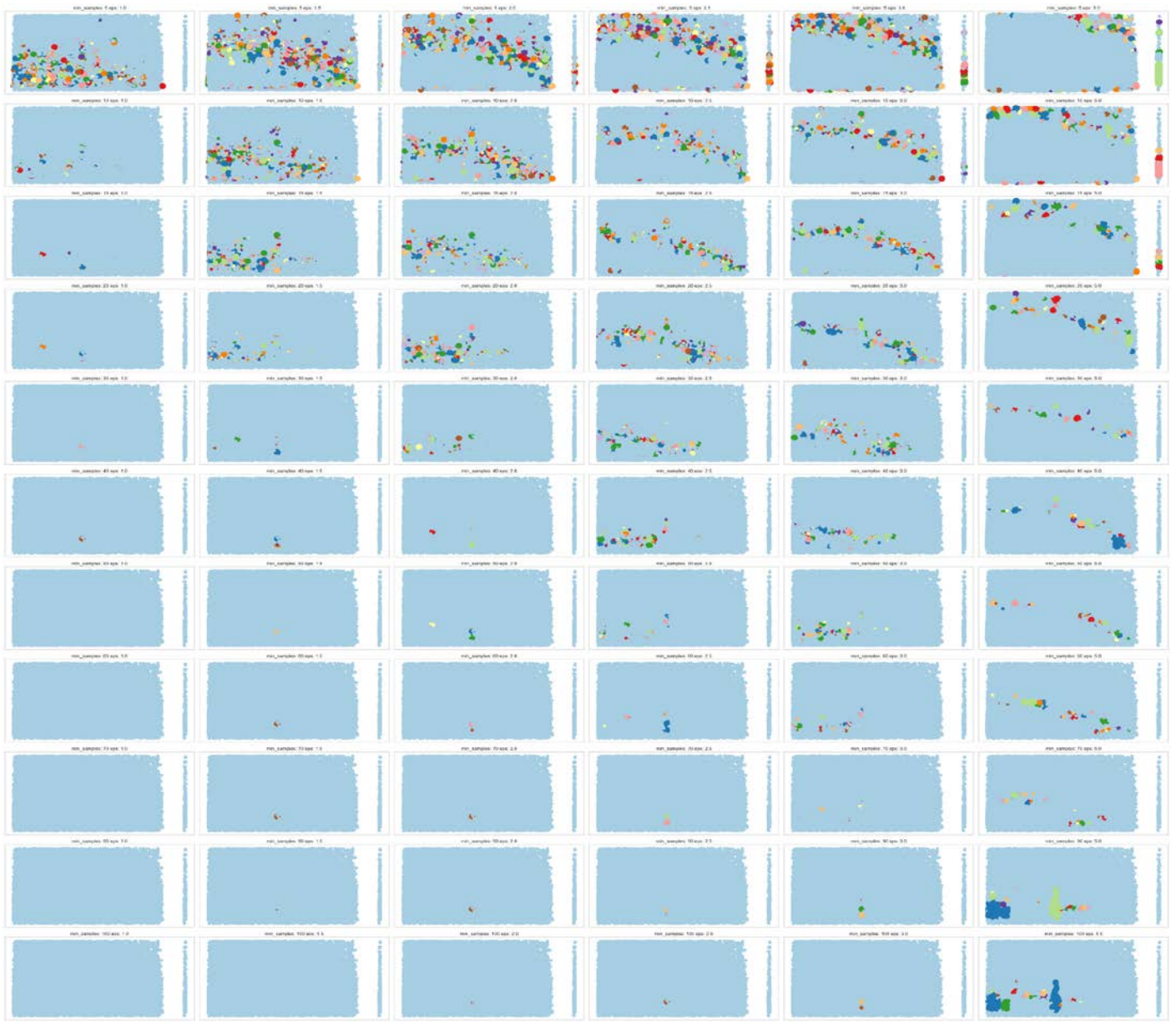


Kmeans clusters characteristics

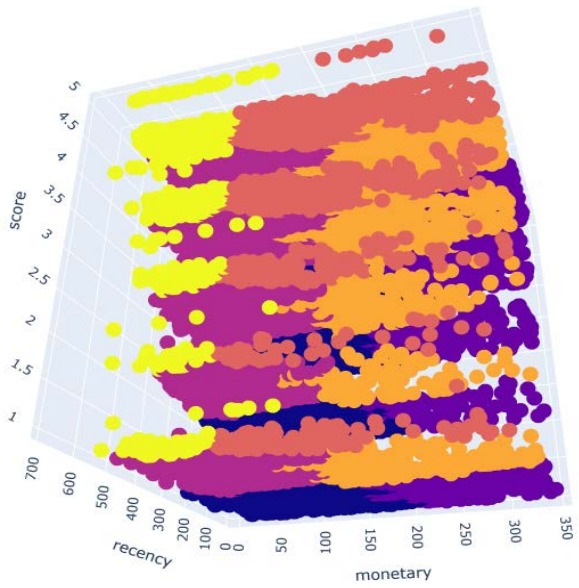
- **cluster 0**: Very good customer. Small amount spender.
- **Cluster 1** : *Weakly regular customers* . Small amount spender.
- **Cluster 2**: good customers. . Small amount spender and regular.
- **Cluster 3**: Non-regular customers. They order less than 1 time each year.
- **Cluster 4**: Fairly loyal good customers but they buy for a high amounts. It can be a customers which buy often the durable goods such as household appliances, furniture ...
- **Cluster 5 :** Reasonably loyal customers. Small amount spender

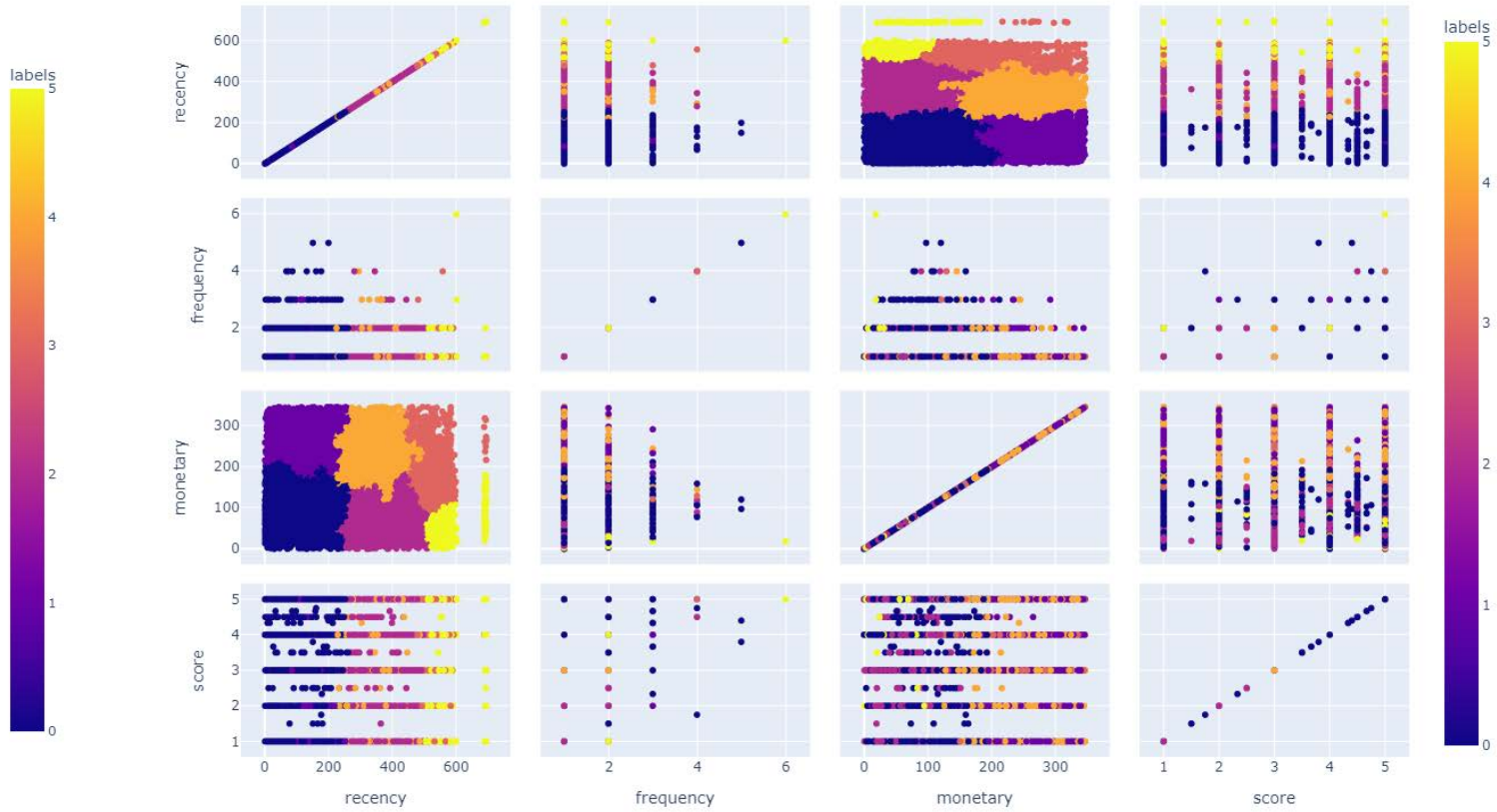
	recency				frequency				monetary				score				cluster_spending
cluster	mean	min	median	max	mean	min	median	max	mean	min	median	max	mean	min	max	count	
0	105.17	1	111.0	204	1.03	1	1.0	7	217.28	136.41	204.44	347.19	4.09	1.0	5.0	9734	2115003.52
1	323.97	254	319.0	421	1.03	1	1.0	4	75.45	0.01	71.82	157.27	4.19	1.0	5.0	16215	1223421.75
2	60.81	1	59.0	124	1.04	1	1.0	14	80.52	0.01	76.24	173.40	4.33	1.0	5.0	17464	1406201.28
3	492.27	396	484.0	695	1.02	1	1.0	6	106.85	0.01	92.68	346.60	4.21	1.0	5.0	13047	1394071.95
4	304.20	204	289.0	498	1.03	1	1.0	4	218.63	139.90	206.92	347.26	4.06	1.0	5.0	8315	1817908.45
5	188.44	125	189.0	258	1.04	1	1.0	8	81.25	0.01	75.17	183.98	4.06	1.0	5.0	19243	1563493.75

DBSCAN

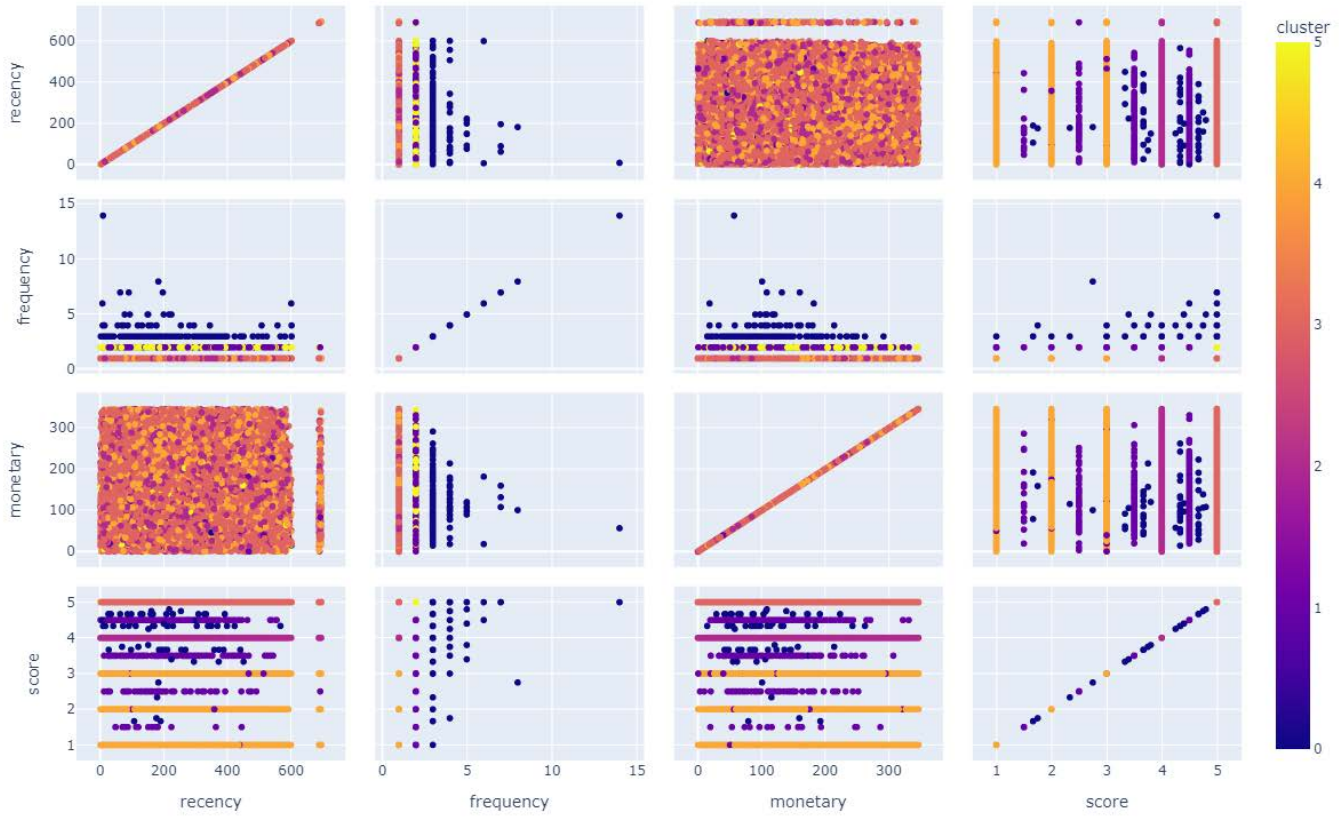


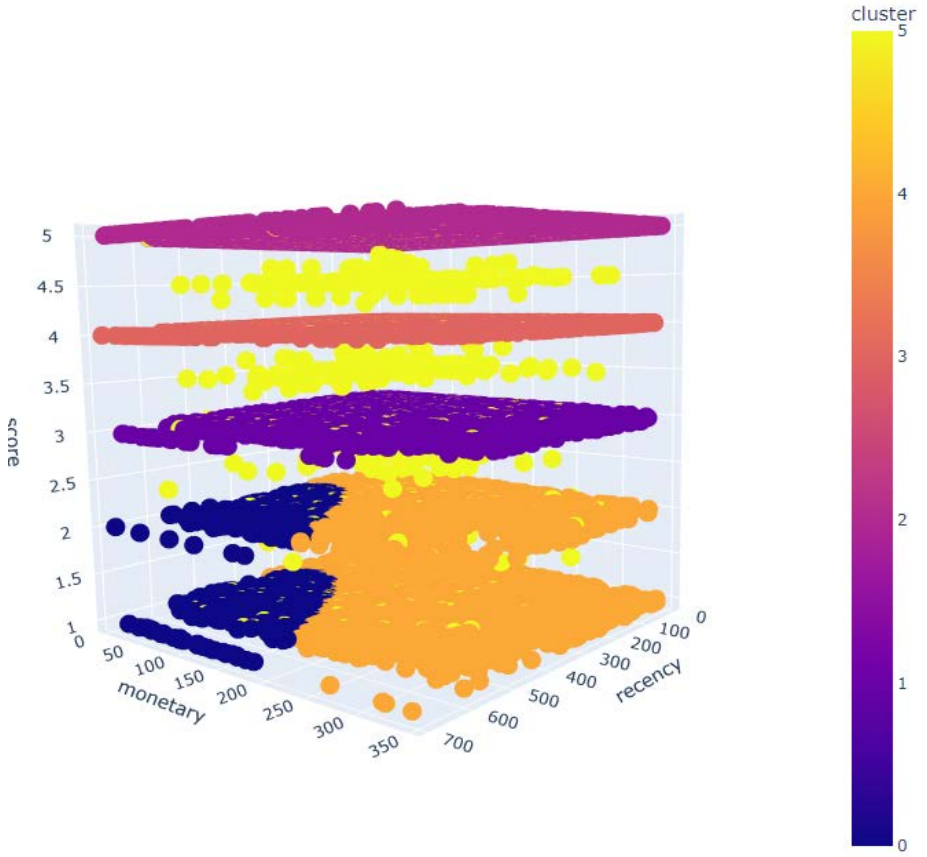
Agglomerative clustering





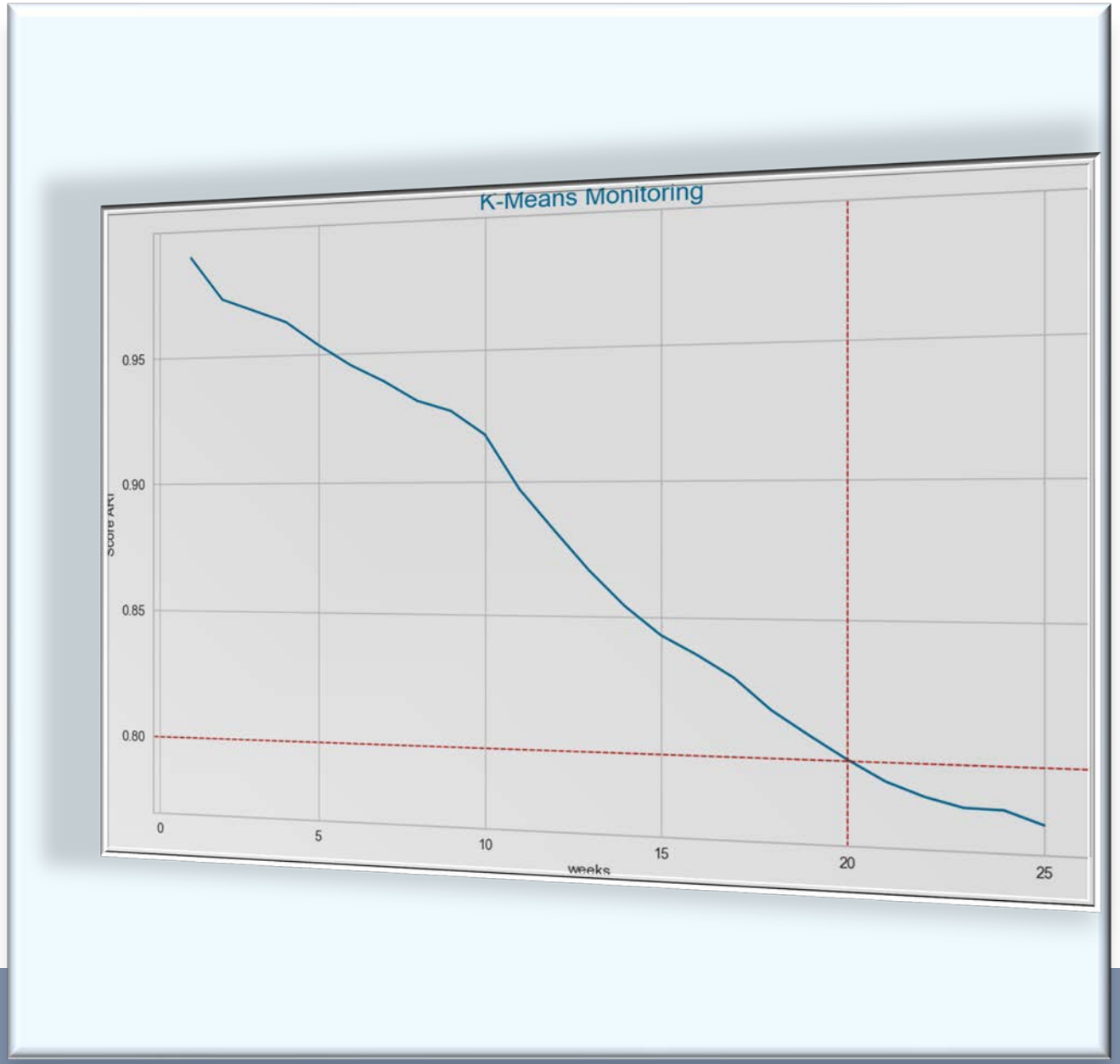
Gaussian Mixture clustering



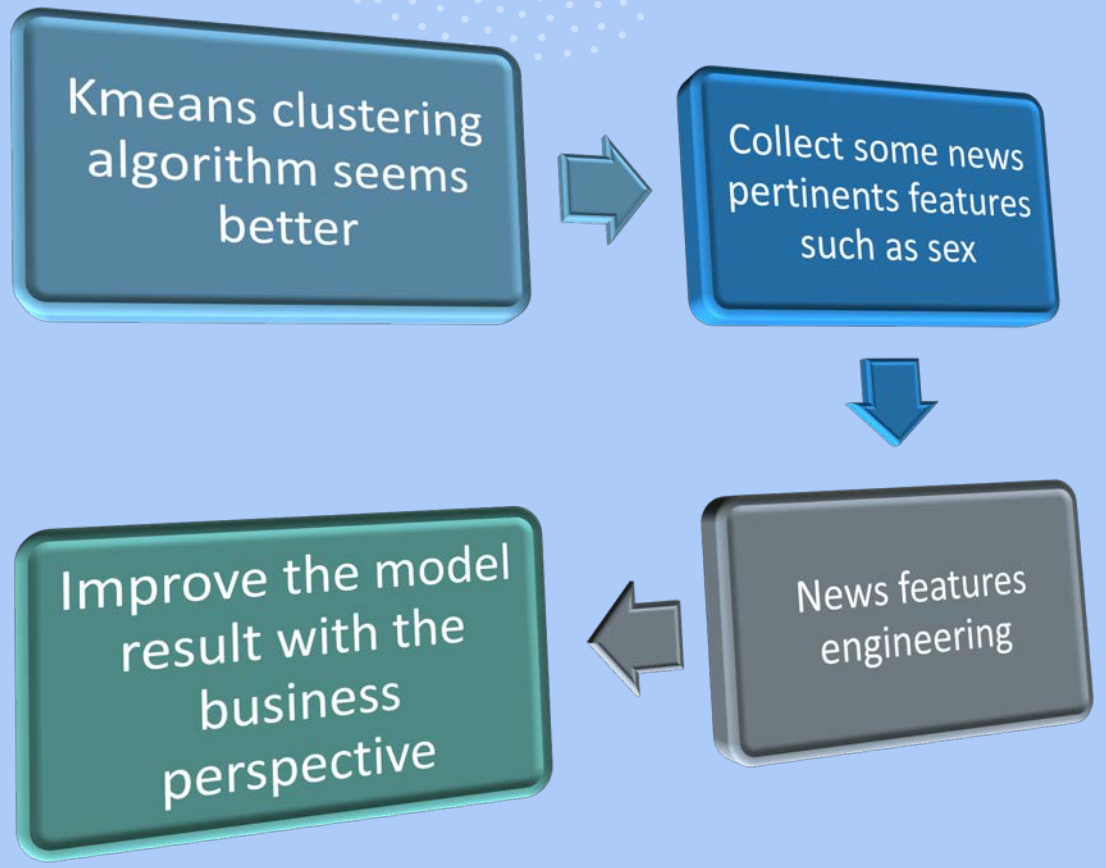


Model Monitoring

- The model will work for 20 then we need to monitor it



Conclusion


Kmeans clustering algorithm seems better
Collect some news pertinents features such as sex
News features engineering
Improve the model result with the business perspective