Olist Ecommerce Analysis

Mengwei Li, Zhenli Min, Mingjun Zhou,
Sahil Sachdev, George Bi, Ke Deng, Yue Sun

Brazilian E-Commerce

Agenda

01 Introduction

02 Exploratory Data Analysis

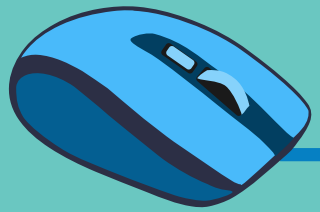
03 Customer Segmentation

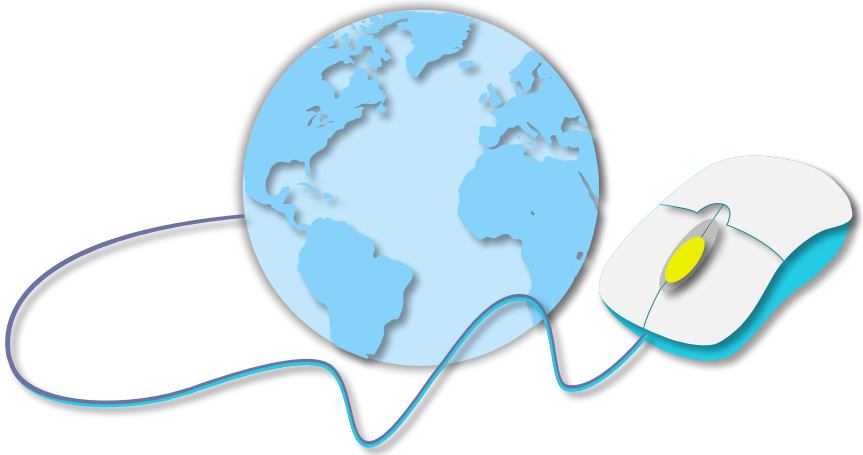
04 Revenue Prediction

05 Review Score Prediction

06 Review Comment Analysis

07 Conclusion





Introduction

Business Case - Purpose

1. Understand Available Data

2. Understand Customers through Segmentation

3. Seller Revenue Prediction

4. Predicting Review Scores

5. Customer Review Analysis

Executive Summary

Data Summary

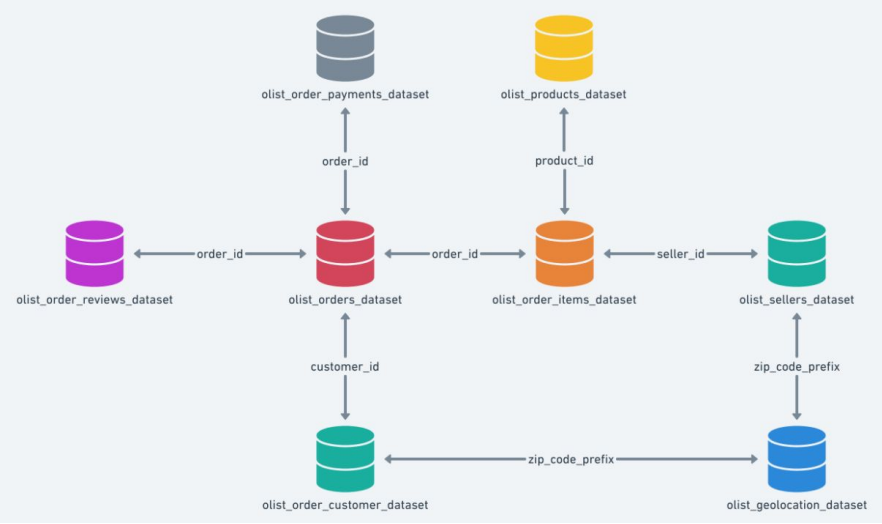
Olist Ecommerce Platform Public Dataset
100,000 orders worth of data from 2016 to 2018
Features Include:
- Order Level Data: Order Id, Order Date, Delivery Date
- Product Level Data: Price, Payment Method, Product Category, Product Reviews
- Customer Level Data: Geolocation, Product Review Score, Product Review Message
- Seller Level Data: Geolocation,
- Marketing Qualified Leads: Lead Category, Catalog Size, Behavior Profile, etc

Overall Approach

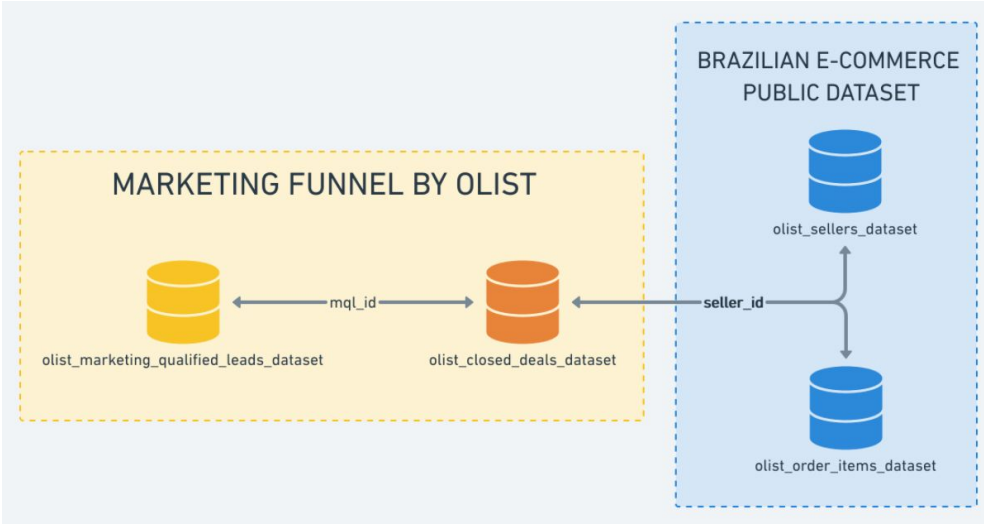
1. Initial Exploratory Data Analysis
2. Data Cleaning and Processing
3. Design and Fit Models depending on Application
4. Insights Extraction

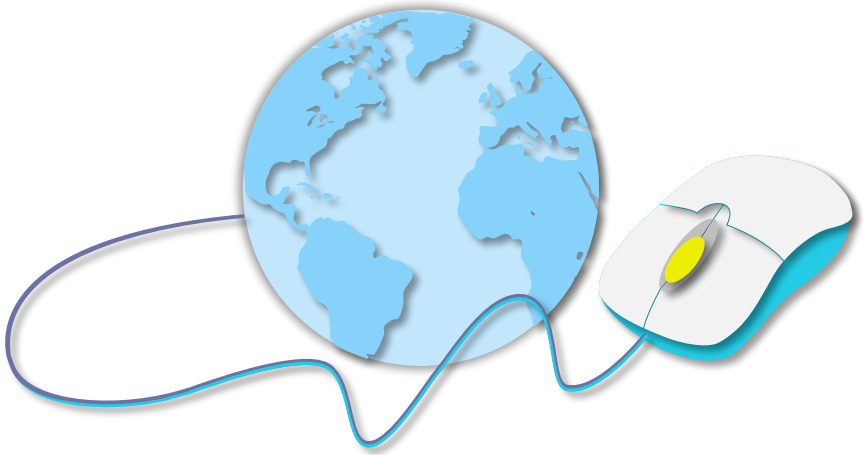
Data Structure and Schemas:

Olist Ecommerce Data Schema



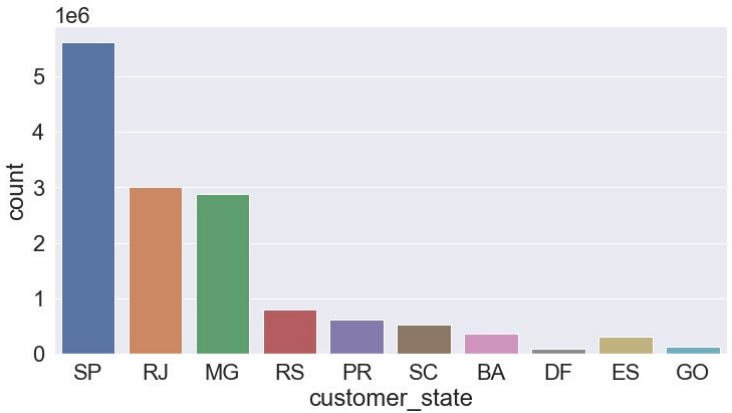
Olist Marketing Funnel Data Schema

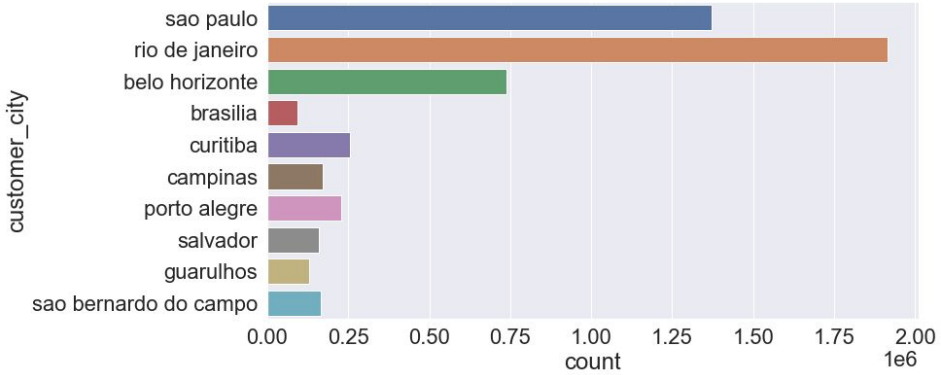


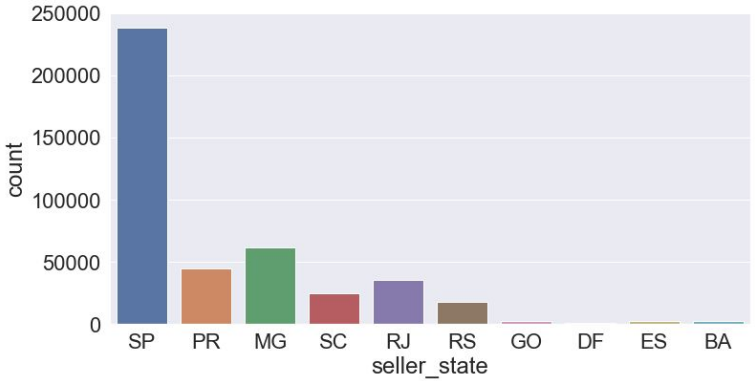


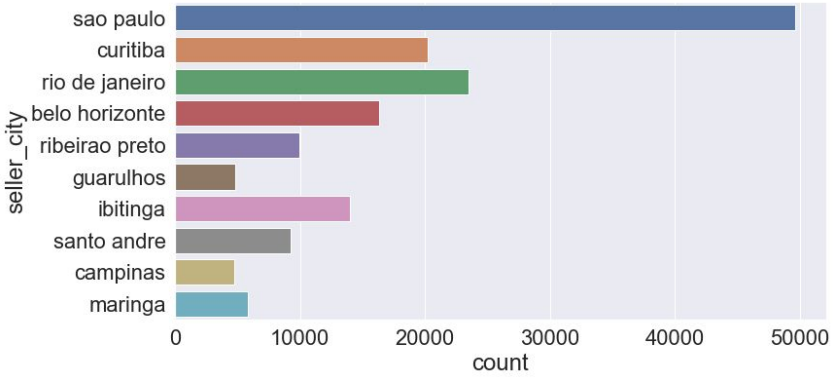
Exploratory Data Analysis

Where's customer & seller?

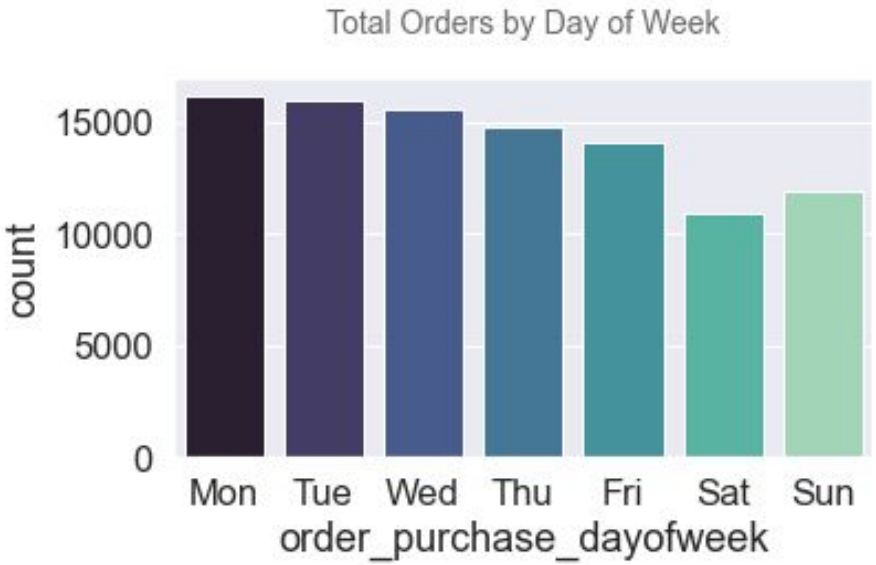


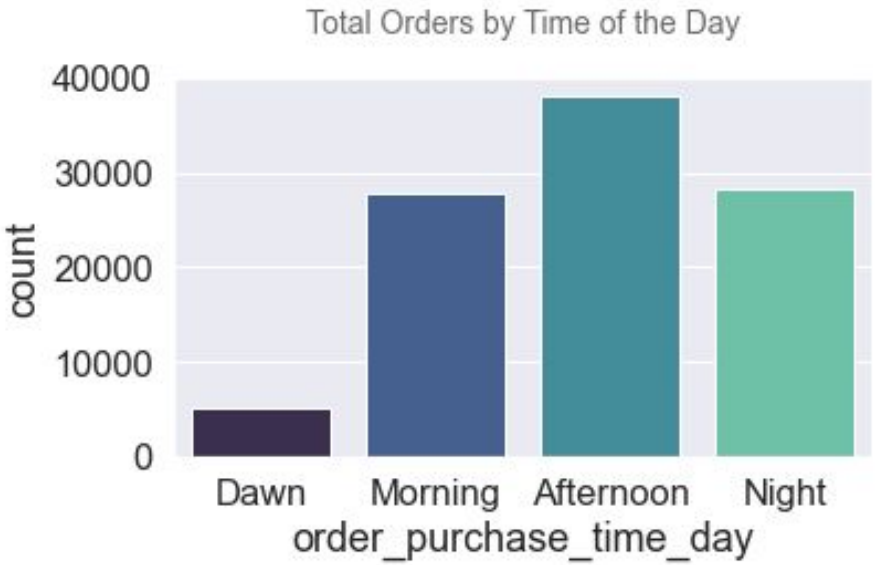




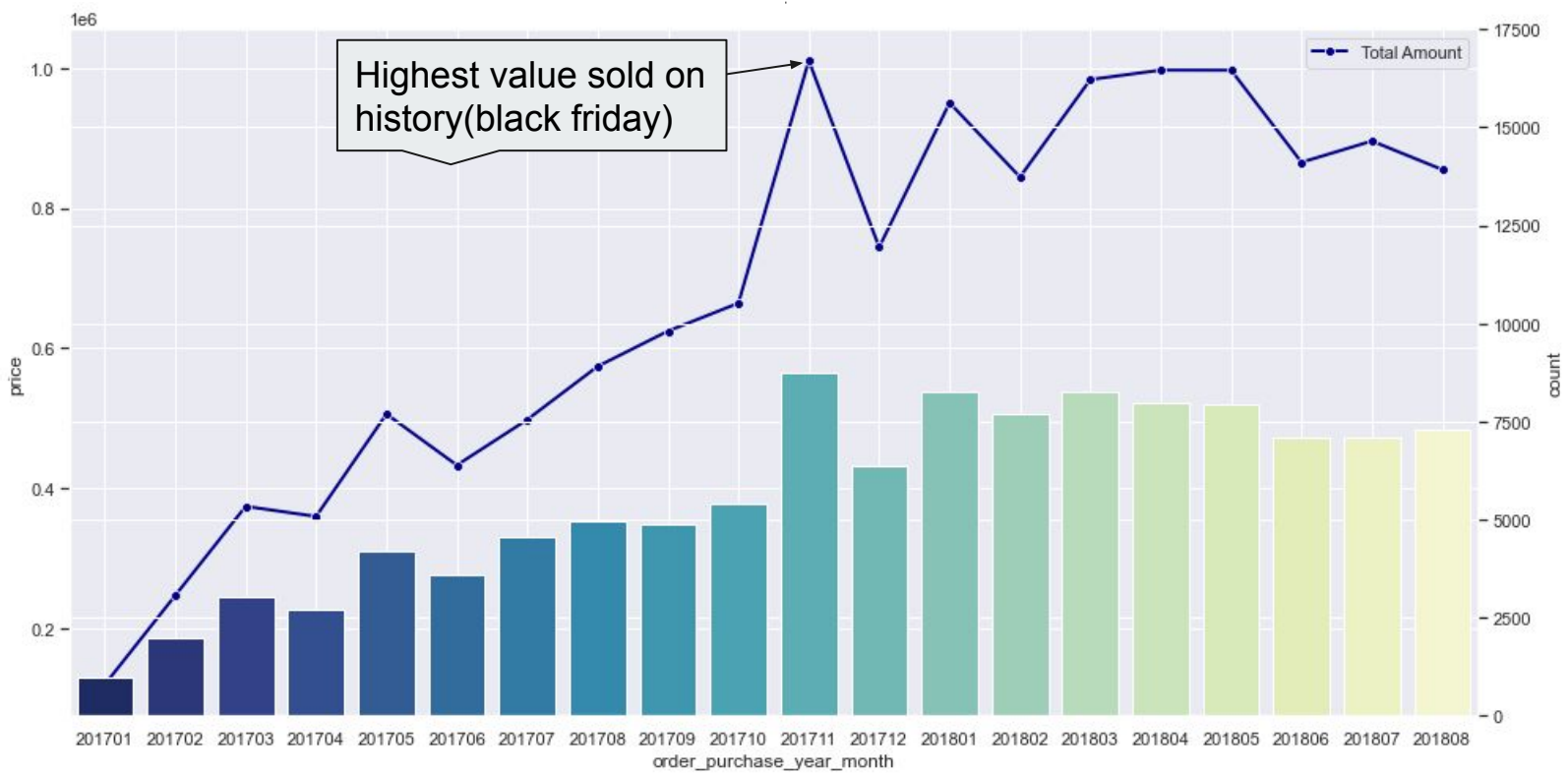


Customer Favorite Day and Time

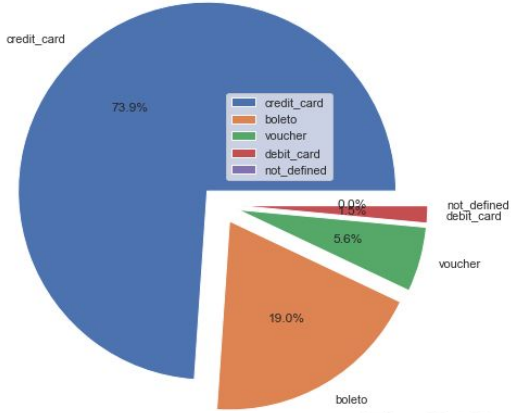


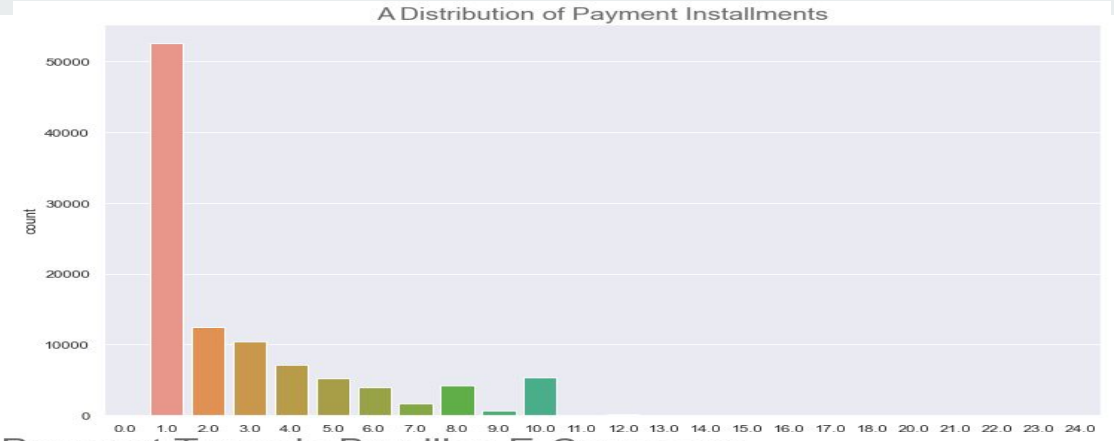


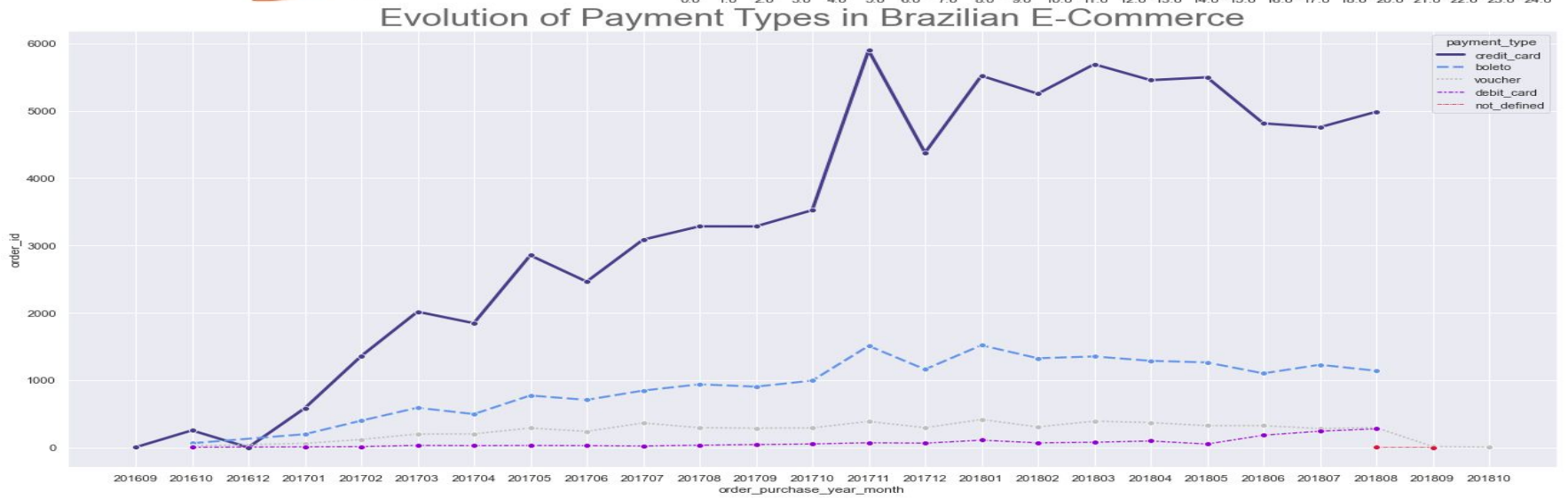
Total Orders and Total Amount Sold



Payments Methods and Installments

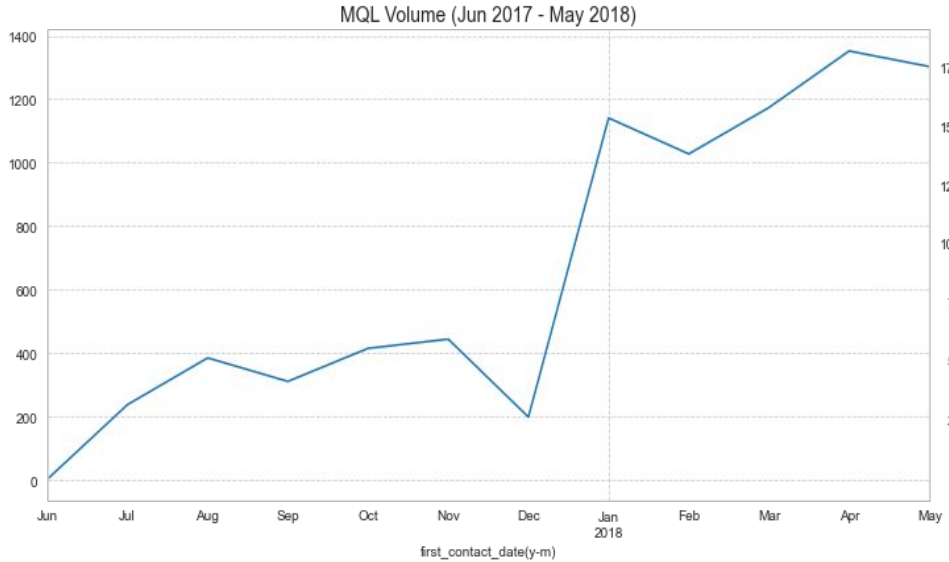


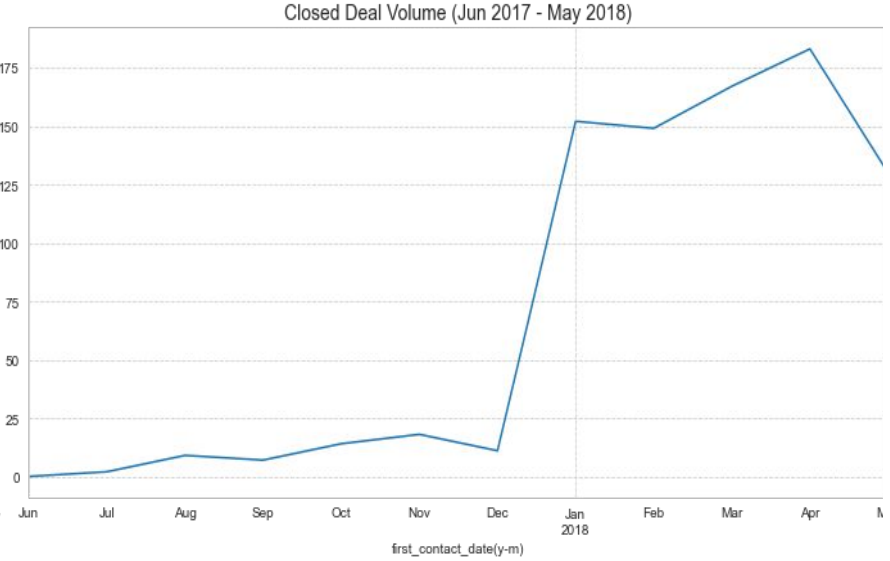




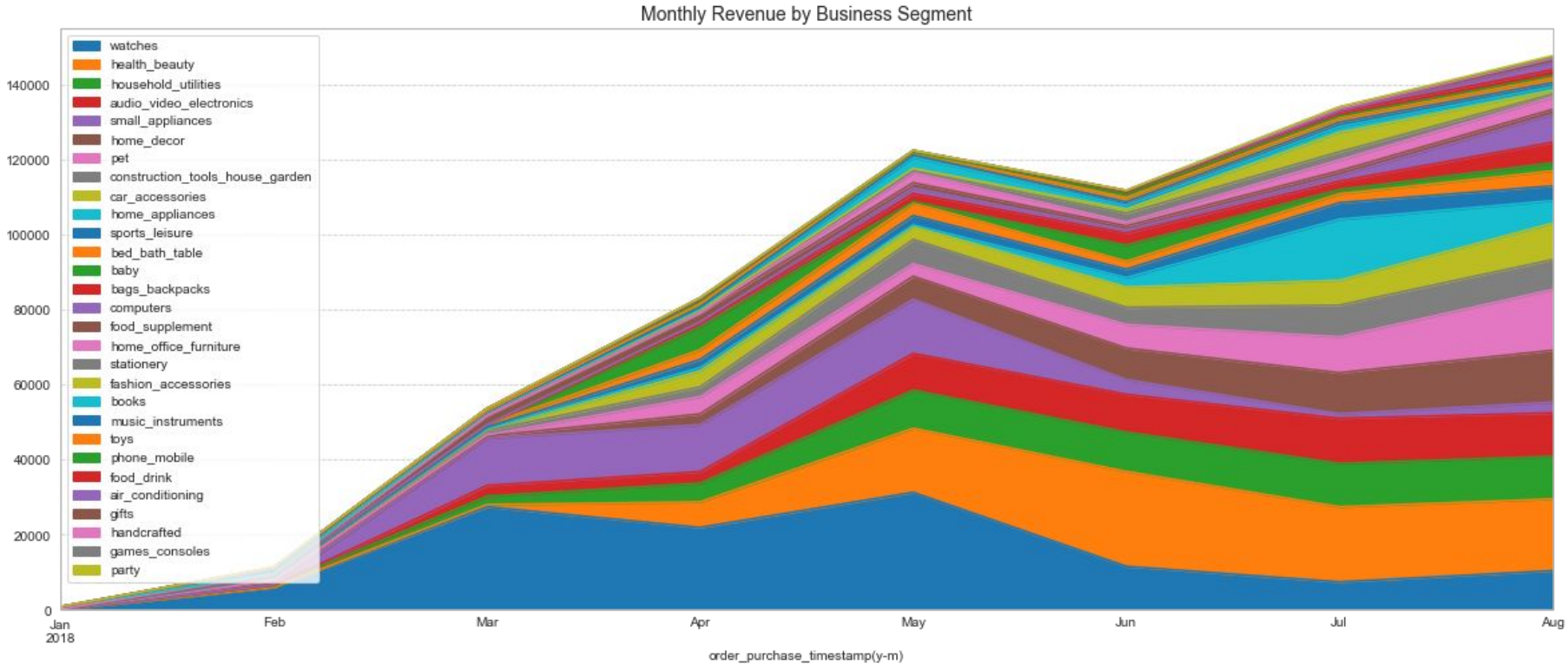
MQL & Closed Deals

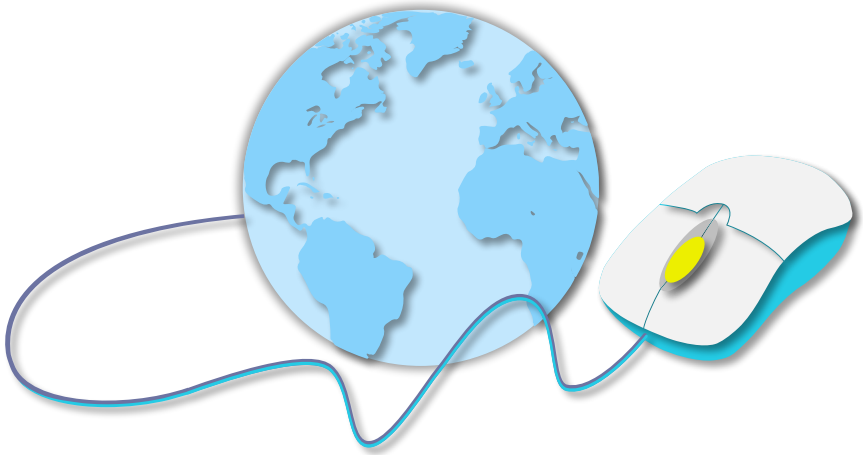
- Often an MQL is a lead who has intentionally engaged with your brand by performing actions like voluntarily submitting contact information, opting into a program, adding e-commerce items to a shopping cart, downloading materials, or repeatedly visiting a website
- A MQL who finally signed up for seller is called a closed deal.





Revenue by business segment

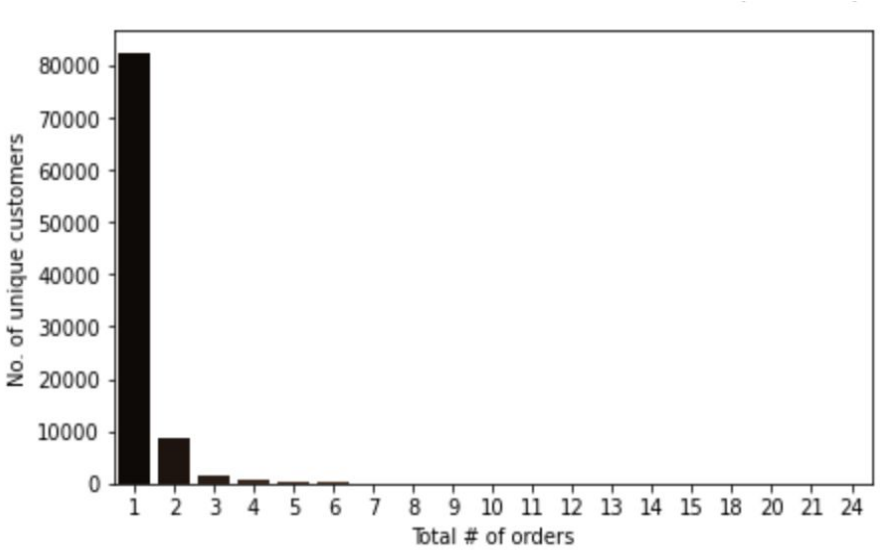




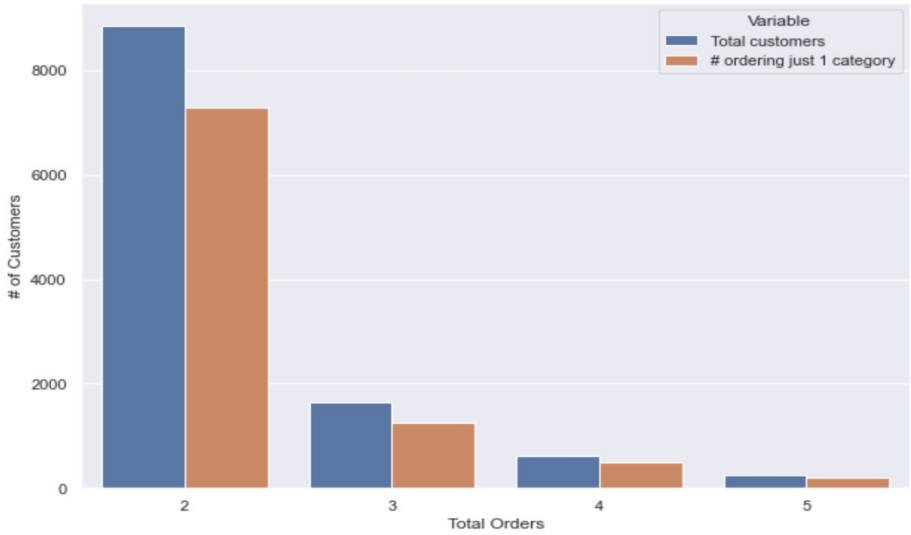
Customer Segmentation

Method 1 - Segmentation by product category

87.6% of customers only ordered once

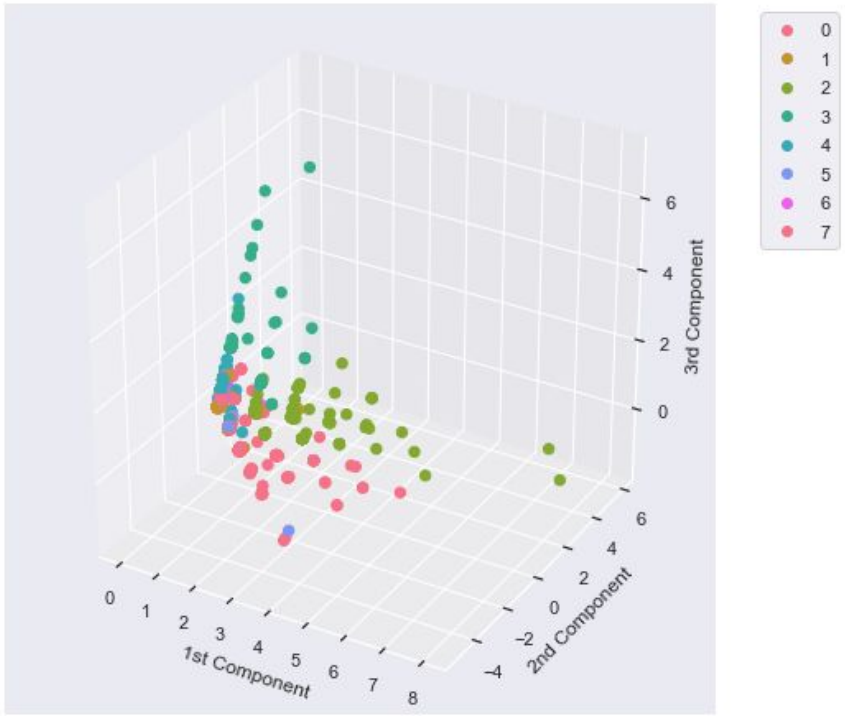


Most customers who purchased more than once (12.4%) stick to only one category

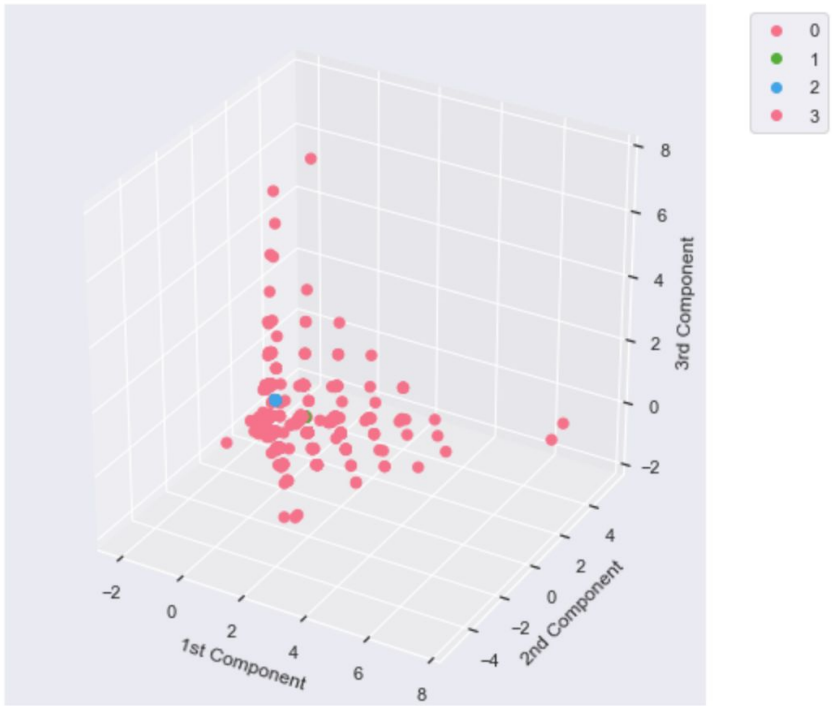


Two examples

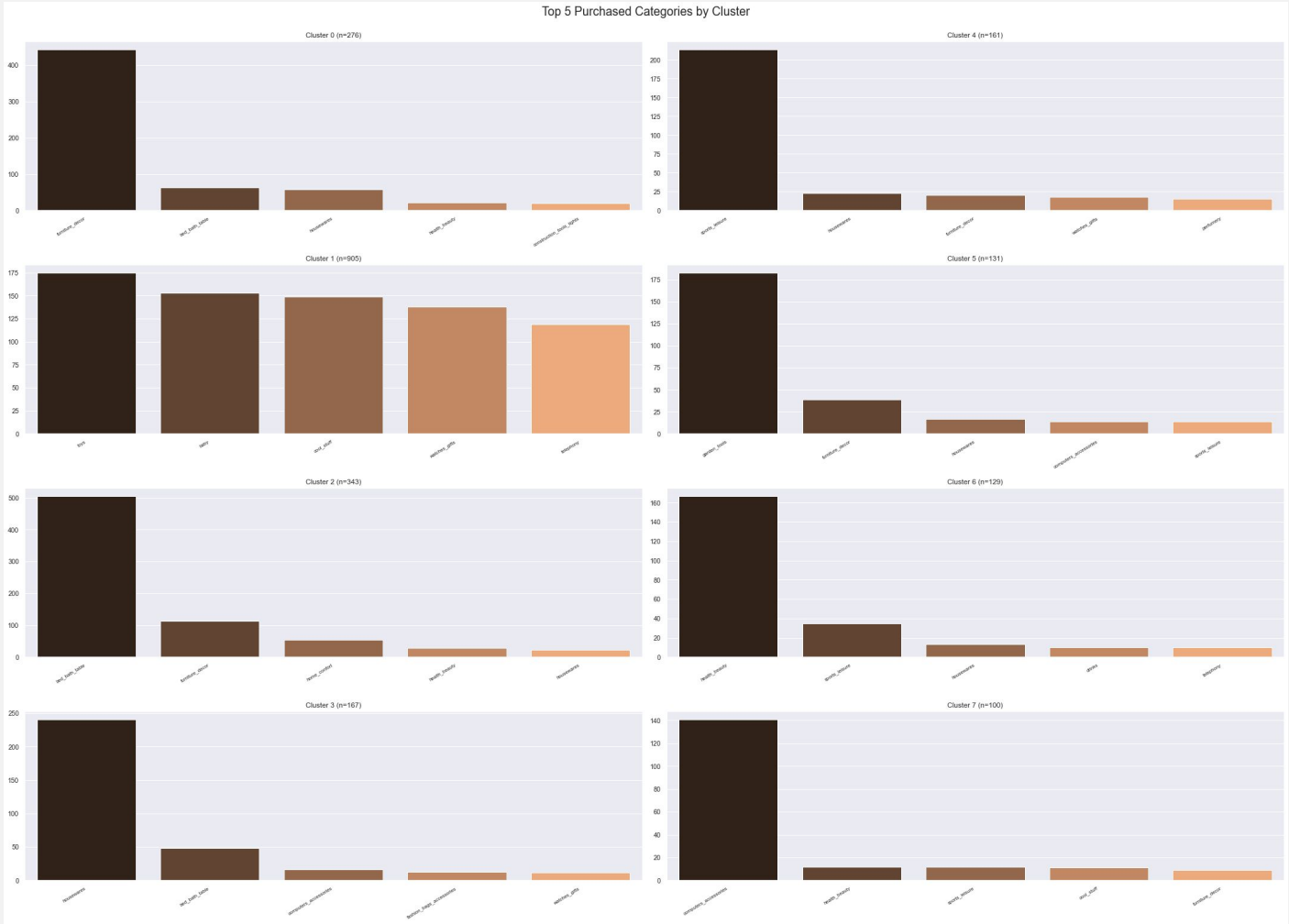
SVD + Agglomerative



PCA + Kmodes

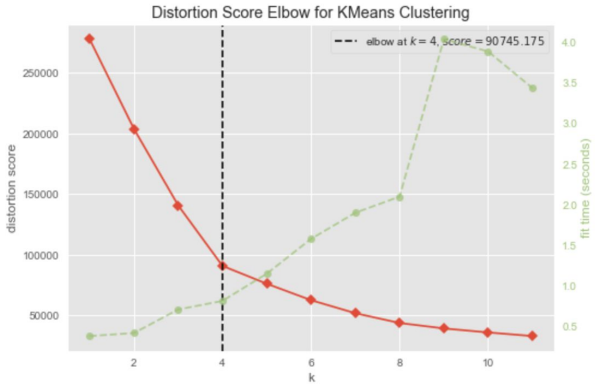


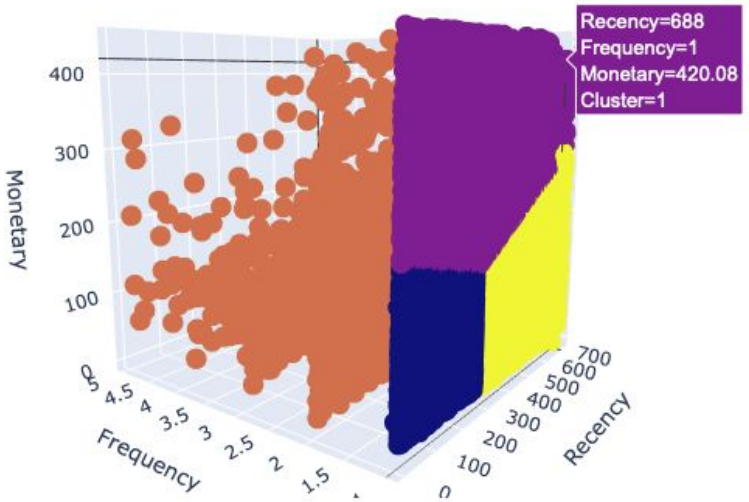
1. furniture and decor
2. toys, baby, cool_stuff, watches_gifts and telephony
3. bed_bath_table
4. housewares
5. sports_leisure
6. garden_tools
7. computers_accessory
8. health_beauty



Method 2.1 - Segmentation by K-Means+RFM

customer_id	Recency	Frequency	Monetary
00012a2ce6f8dcda20d059ce98491703	299	1	114.74
000161a058600d5901f007fab4c27140	420	1	67.41
0001fd6190edaaf884bcaf3d49edf079	560	1	195.42
0002414f95344307404f0ace7a26f1d5	389	1	179.35
000379cdec625522490c315e70c7a9fb	161	1	107.01
...	...	...	...
fffcb937e9dd47a13f05ecb8290f4d3e	179	1	91.91
fffecc9f79fd8c764f843e9951b11341	165	3	81.36
fffeda5b6d849fbd39689bb92087f431	110	1	63.13
ffff42319e9b2d713724ae527742af25	89	1	214.13
ffffa3172527f765de70084a7e53aae8	368	1	45.50

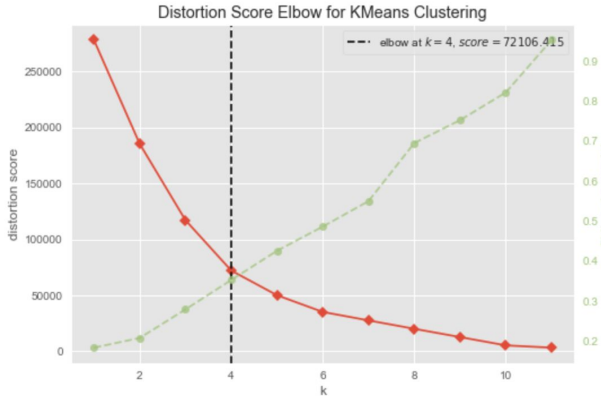


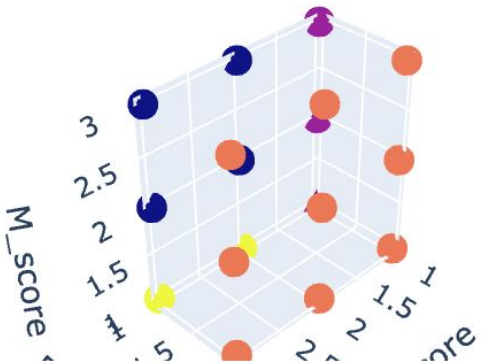


Cluster 1: Rookies - Our Newest Customers. First time buyers.(group size: 43417)
Cluster 2: Whales - Our Highest Paying Customers. (group size: 15521)
Cluster 3: Slipping - Once Loyal, Now Gone. (group size: 31091)
Cluster 4: Promising- Faithful customers (group size: 2649)

Method 2.2 - Segmentation by K-Means+RFM_Score

customer_id	R_score	F_score	M_score
00012a2ce6f8dcda20d059ce98491703	2	1	2
000161a058600d5901f007fab4c27140	1	1	1
0001fd6190edaaf884bcaf3d49edf079	1	1	3
0002414f95344307404f0ace7a26f1d5	1	1	3
000379cdec625522490c315e70c7a9fb	3	1	2
...	...	...	...
fffcb937e9dd47a13f05ecb8290f4d3e	2	1	2
fffecc9f79fd8c764f843e9951b11341	2	2	2
fffeda5b6d849fbd39689bb92087f431	3	1	1
ffff42319e9b2d713724ae527742af25	3	1	3
ffffa3172527f765de70084a7e53aae8	1	1	1





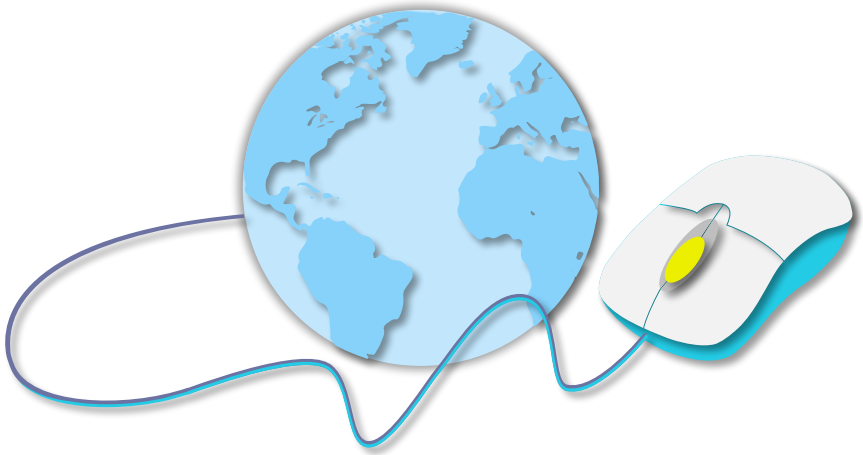
Recency: 1-3
Frequency: 1-2
Monetary: 1-3
(1: lowest / 3: highest)

Cluster 1: Whales - Our Highest Paying Customers. (group size: 40269)
Cluster 2: Slipping - Once Loyal, Now Gone. (group size: 29683)
Cluster 3: Promising - Faithful customers. (group size: 2649)
Cluster 4: Rookies - Our Newest Customers. (group size: 20077)

Method 2.3 - Segmentation by combination of RFM_Score

customer_id	R_score	F_score	M_score	RFM_score
00012a2ce6f8dcda20d059ce98491703	2	1	2	212
000161a058600d5901f007fab4c27140	1	1	1	111
0001fd6190edaaf884bcaf3d49edf079	1	1	3	113
0002414f95344307404f0ace7a26f1d5	1	1	3	113
000379cdec625522490c315e70c7a9fb	3	1	2	312
...	...	...	...	...
fffcb937e9dd47a13f05ecb8290f4d3e	2	1	2	212
fffecc9f79fd8c764f843e9951b11341	2	2	2	222
fffeda5b6d849fbd39689bb92087f431	3	1	1	311
ffff42319e9b2d713724ae527742af25	3	1	3	313
ffffa3172527f765de70084a7e53aae8	1	1	1	111

Cluster 1: Core - Best Customers (RFM_Score: 323)
Cluster 2: Loyal - Most Loyal Customers (RFM_Score: 221/222/121/122)
Cluster 3: Rookies- Newest Customers(RFM_Score: 31X)
Cluster 4: Whales - Highest Paying Customers (RFM_Score: 1X3/2X3)
Cluster 5: Slipping - Once Loyal, Now Gone. (RFM_Score: 11X)
Cluster 6: Regular- Average in R, F and M. (the other conditions)



Seller Revenue Prediction

Seller Revenue Prediction - Overview

01

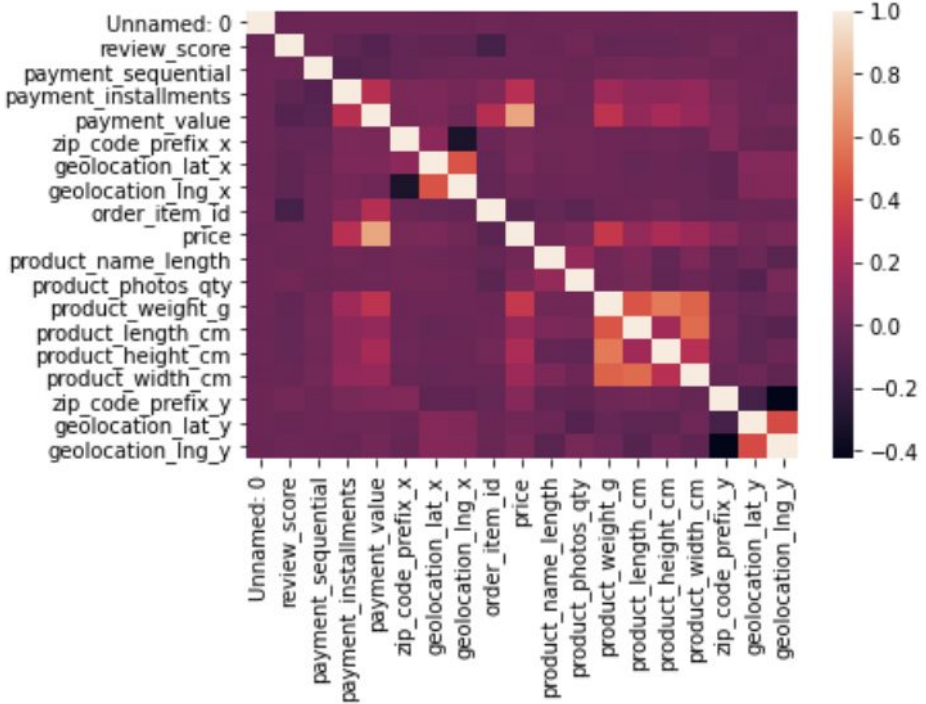
Data Cleaning

Dropping Features w > 70% NA Values
Dropping Rows w NA Values

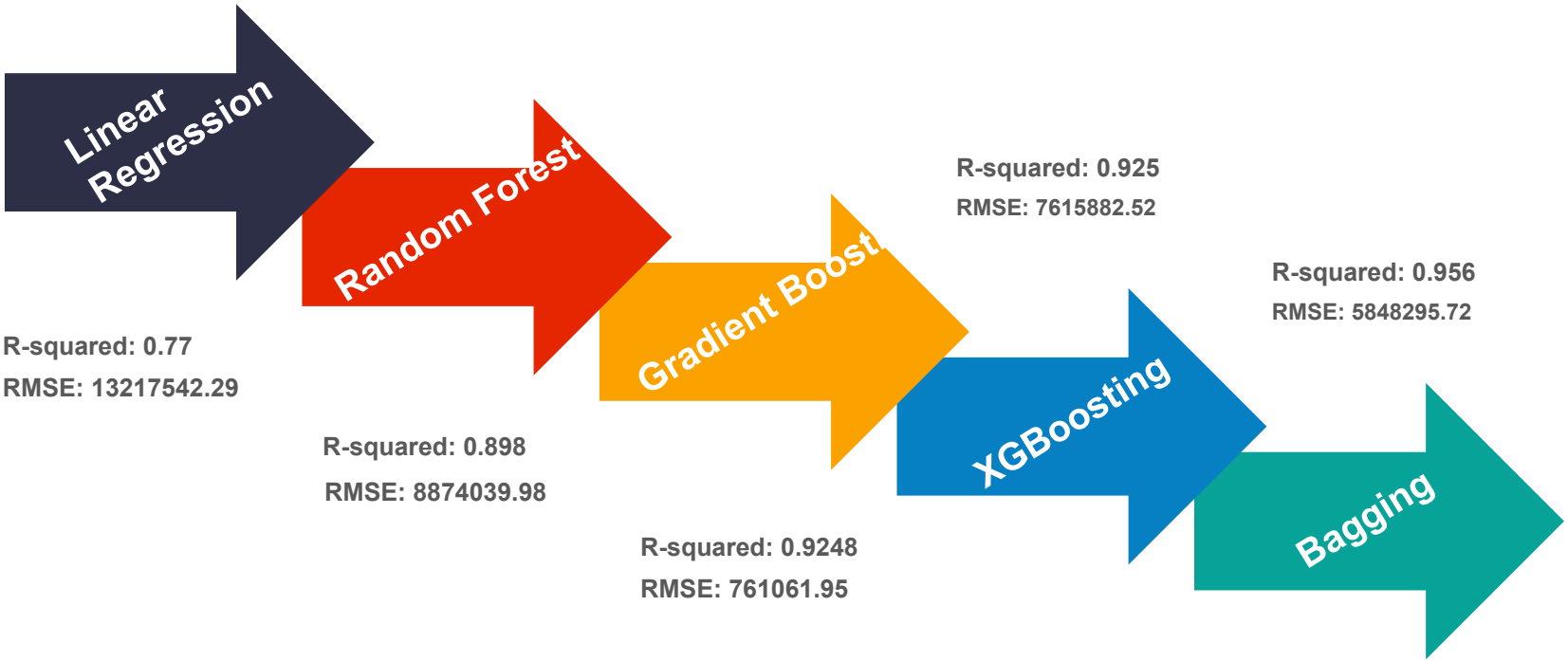
02

Features and Correlations

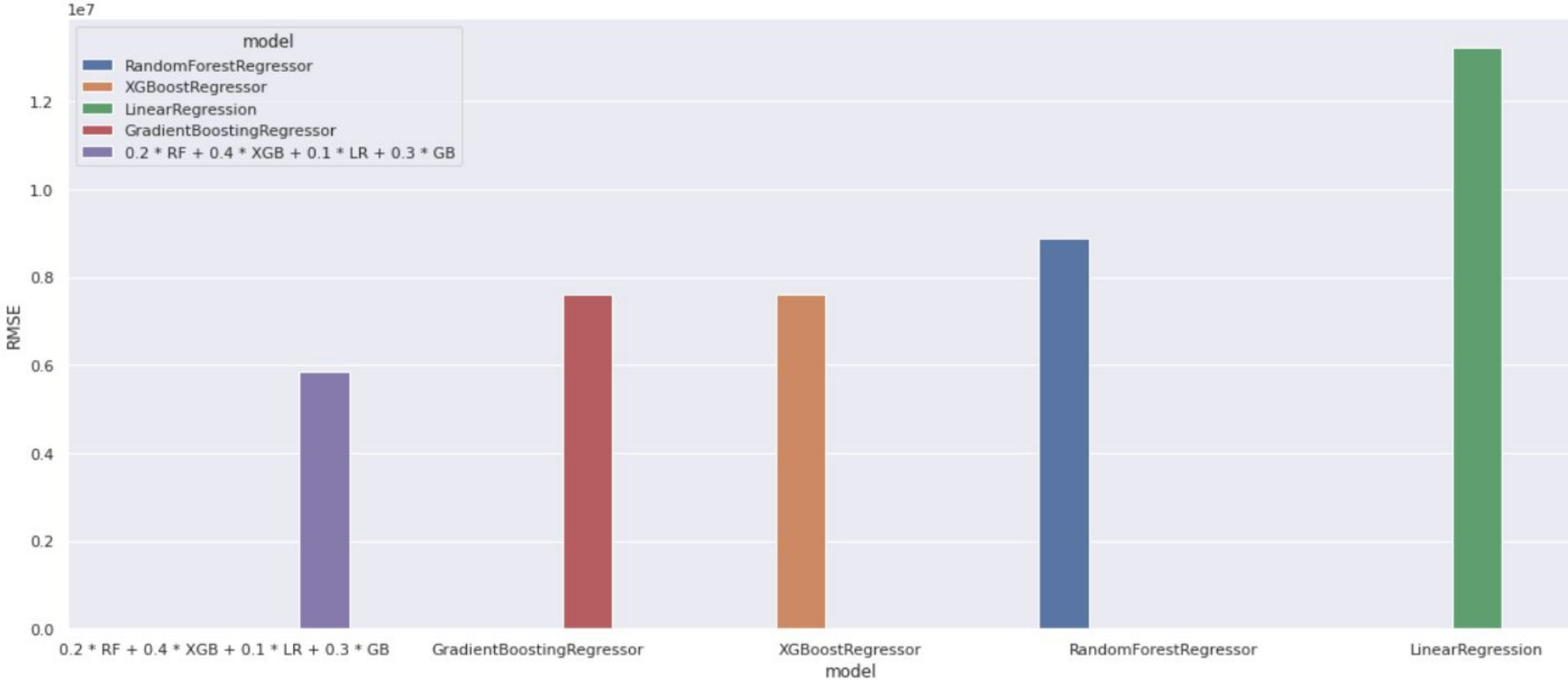
Dropping Features w > 70% NA Values
Dropping Rows w NA Values
Creating TotalPaymentValue feature



3. Seller Revenue Prediction Models

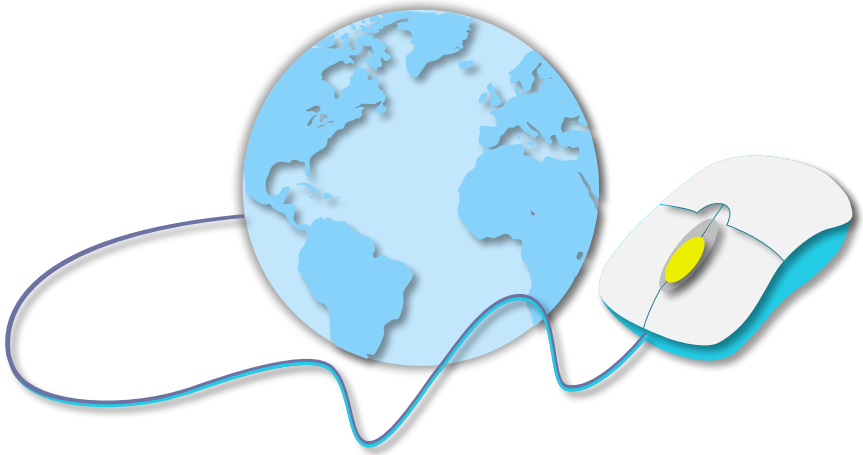


4. Model Performance



5. Model Prediction (Example)

seller_id	test_data	bagging_result
30829de4562ab	341994.48	311443.94
e6218512d16fca	7281.00	10171.79
c3867b4666c7d	10820785.60	13597808.01
8629c241b3662	15179.58	15079.89
c7fdb77fdbff3c	18061.80	10198.72
0ff83046c3fa22	11743.20	14319.86
aae5e7b457a3	6522.60	7767.42
5d76d80a2f5f3	8330.40	9179.47
b5abfd436adc	18168.80	17095.05
a2b5b6105ea59	52587.04	58028.61



Review Score Prediction

ML Model for Review Score

Order of operations:

1.Data Cleaning
2 Merge different dataset
3.Fit model on train
4.Apply model to test
5.Classification report

Reason to use Review Score ML Model

1 Using ML model to predict review scores to build better recommendation advice when shopping online.
2 find out reason behind negative rating for further improvement.

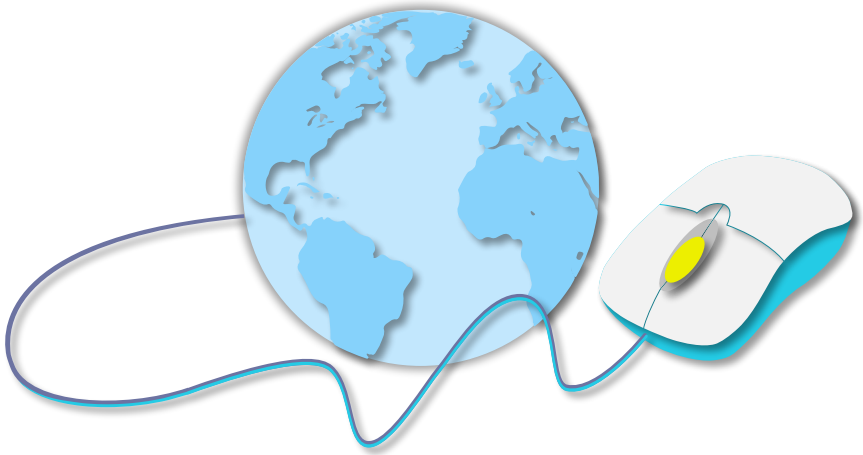
Model Results

Model	RMSE	Train Accuracy	Test Accuracy
Random Forest	0.424	0.8291	0.8271
Logistic Regression	0.3735	0.8572	0.8605
Decision Tree	0.516578	0.7468	0.7331
GBDT	0.367503	0.874	0.8649

Advantage of GBDT:

- superior ability to find nonlinear interactions automatically.
- cross validation in each iteration
- handle missing value

GBDT Model performs best both in RMSE Score and Accuracy Score



Review Comment Analysis

POSITIVE REVIEW COMMENT

- Reliable seller, product ok and delivery before deadline.
- I got exactly what I expected. The other sellers' orders were delayed, but this one arrived on time.
- I'm completely in love, super responsible, reliable store!

NEGATIVE REVIEW COMMENT

- I'd like to know what's been going on, but I always got it back, and that's what's going on now.
- Very inferior product, badly finished.
- "I made my purchase thirty days ago and I haven't received my product yet.
  You need better deliveries."


freight
still
better
expected
paid
will
beautiful
took
well packaged
much
according
delivered within
deadline
great
far
answer
send
wrong
photo
today
company
reached
solve
waiting
one
original
advertisement
within
thank
problem
give
always
bought two
mail
complaint
size
little
fit
well packed
information
never
delivered
excellent
perfect
box
buy
exactly
invoice
deliver
without
merchandise
post office
come
price
work
arrived well
made
receive
seller
delivery time
material
arrived
unfortunately
quality
delivery
everything
arrive
two
ok
gave
canceled
satisfied
speed
quality product
contact
super
long
another
baratheon
ordered
easy
lack
see
note
bought
good
change
well
need
lot
worth
month
take
super recommend
term
perfect condition
got
say
announced
super fast
piece
product delivered
right
expectation
loved
correct
lannister store
fast
missing
correctly
received
good product
email
black
exchange
appear
order
nothing
stark
thought
congratulations
even
working
product came
different
product yet
liked
recommend
arrived
make
store
purchase
great quality
look
now
evaluate
day
want
packaging
request
fast delivery
way
really
combined
lannister
requested
purchased
good quality
really liked
value
correct product
stipulated deadline
best
product
expected deadline
disappointed
description
think
cover
know
partner
found
curtain
bad
site
asked
item
small
model
port
use
return
came
excellent product
time
color
inside
address
defect
customer
yet
kit
used
hope
great product
already
targaryen
delay
broken
described
week
excellent quality
sent
clock
unit
date
service

	precision	recall	f1-score	support
0	0.78	0.81	0.80	2902
1	0.92	0.90	0.91	7074
accuracy			0.88	9976
macro avg	0.85	0.86	0.85	9976
weighted avg	0.88	0.88	0.88	9976

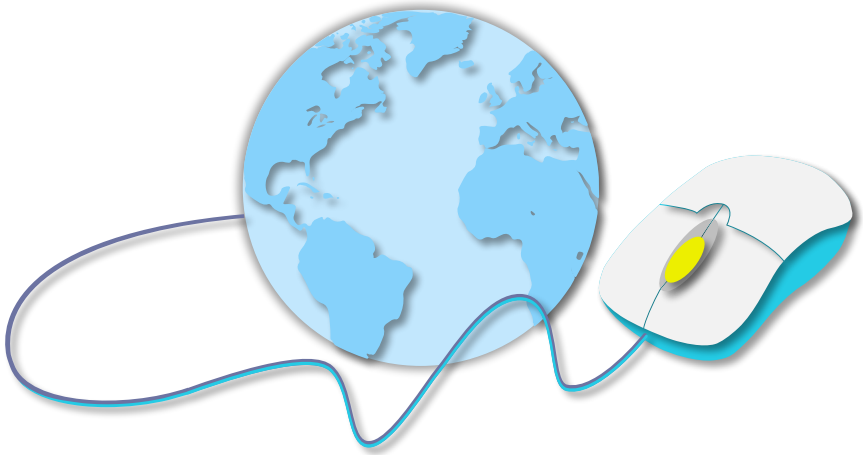
Binomial Naive Bayes

Logistic Regression

	precision	recall	f1-score	support
0	0.82	0.85	0.83	2902
1	0.94	0.92	0.93	7074
accuracy			0.90	9976
macro avg	0.88	0.89	0.88	9976
weighted avg	0.90	0.90	0.90	9976

	precision	recall	f1-score	support
0	0.76	0.77	0.77	2902
1	0.90	0.90	0.90	7074
accuracy			0.86	9976
macro avg	0.83	0.83	0.83	9976
weighted avg	0.86	0.86	0.86	9976

Ada Boost

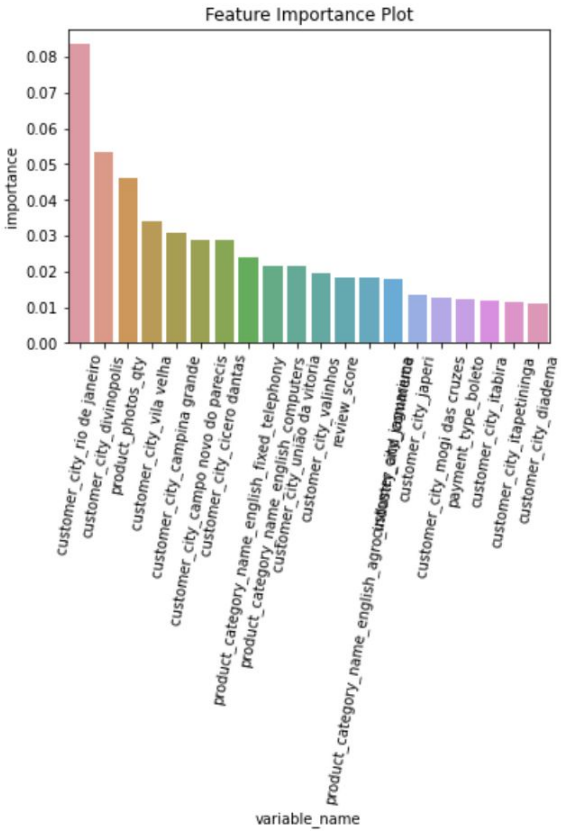


Future Analysis

Suggested Future Analysis - Customer Revenue Prediction per Order for LTV

For Future Analysis:

- We can conduct Customer Revenue Prediction based on Order Data
- This can be built upon Coco's Customer Segmentation that she shared above.
- Using the RFM clusters, we can predict customer order revenue
    - This will help the businesses understand which customers are the most profitable and desired for targeting



Customer Segments by RFM

Cluster 1: Core - Best Customers (RFM_Score: 323)
Cluster 2: Loyal - Most Loyal Customers (RFM_Score: 221/222/121/122)
Cluster 3: Rookies- Newest Customers(RFM_Score: 31X)
Cluster 4: Whales - Highest Paying Customers (RFM_Score: 1X3/2X3)
Cluster 5: Slipping - Once Loyal, Now Gone. (RFM_Score: 11X)
Cluster 6: Regular- Average in R, F and M. (the other conditions)

Q&A

Thank you!