DA NANG UNIVERSITY OF ECONOMICS
FACULTY OF E-COMMERCE

46K29.1

KHDL & PTKD

Capstone 1

Olist Customer Report

GVHD

Trương Hồng Tuấn

Thành viên

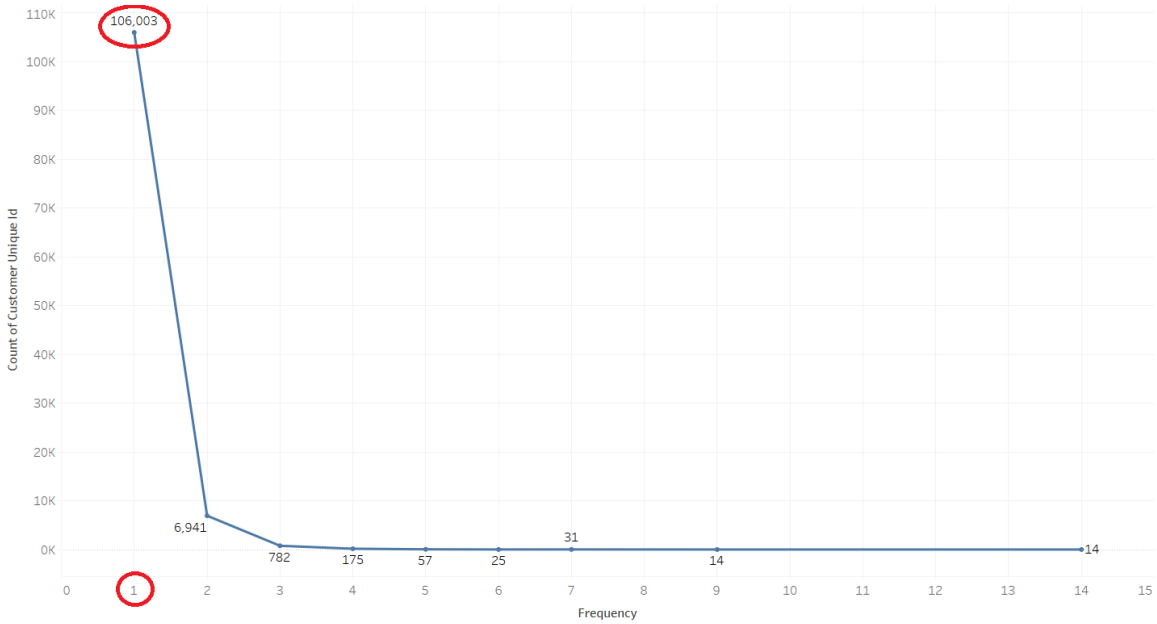
Nguyễn Thị Ngọc Xuân
Phan Ngọc Tiến

MỤC LỤC

I. Vấn đề doanh nghiệp 2
II. Mô tả dữ liệu 3
III. Tiền xử lý 6
IV. Phân tích 7
1. Phân tích EDA: Tổng quan tình hình KH 7
2. Customer Sentiment (count word) 13
3. Customer Segmentation (RFM) 20
Link tableau: 27

I. Vấn đề doanh nghiệp

Cửa hàng Olist là một doanh nghiệp thương mại điện tử có trụ sở tại Sao Paulo, Brazil. Đây là trung gian kết nối các doanh nghiệp nhỏ khác nhau với những khách hàng muốn mua sản phẩm của họ. Cửa hàng thu hút số lượng lớn khách hàng tham gia, hoạt động mạnh mẽ, tuy nhiên lượng khách hàng quay trở lại mua lần 2 rất ít. Do đó, doanh nghiệp cần tận dụng những dữ liệu sẵn có để giải quyết vấn đề trên, tìm cách tăng sự hài lòng, tăng lòng trung thành của khách hàng.



Hình 1: Tần suất mua hàng của khách hàng ở Olist Store

Các loại phân tích áp dụng để giải quyết vấn đề của doanh nghiệp:

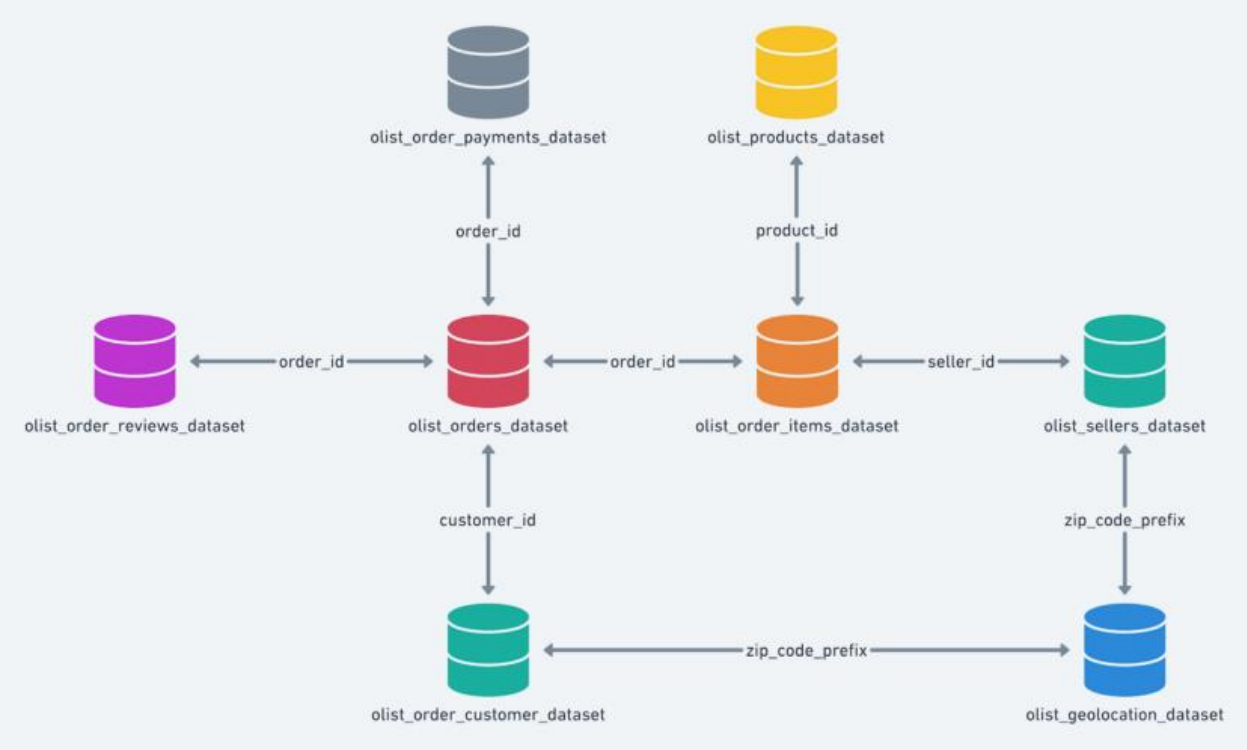
- Phân tích EDA
- Phân tích cảm xúc khách hàng dựa trên review (count word)
- Customer Segmentation (RFM)

II. Mô tả dữ liệu

Bộ dữ liệu được lấy từ Kaggle chứa thông thi khoảng 100.000 đơn đặt hàng gồm 9 datasets được thực hiện tại nhiều thị trường trong khoảng thời gian từ 2016 đến năm 2018.

Dữ liệu của Olist (9 files csv):

1. olist_customers_dataset.csv (data)
2. olist_geolocation_dataset.csv(geo_data)
3. olist_order_items_dataset.csv(order_itemdata)
4. olist_order_payments_dataset.csv(pay_data)
5. olist_order_reviews_dataset.csv(rev_data)
6. olist_orders_dataset.csv(orders)
7. olist_products_dataset.csv(order_prddata)
8. olist_sellers_dataset.csv(order_selldata)
9. product_category_name_translation.csv(order_prd_catdata)



Mô tả cơ bản dữ liệu (một số cột tiêu biểu):

Columns Name	Data Type, Length	Description	Present
customer_id	object	Key to the orders dataset. Each order has a unique customer_id.	Unique customer id to merge with orders dataset
customer_unique_id	object	unique identifier of a customer.	represent customer unique id
customer_zip_code_prefix	int64	first five digits of customer zip code	represent first five zip code of geolocation customer
customer_city	object	customer city name	a customer city name in brazil
customer_state	object	customer state	represent 2 letters of name customer sate
geolocation_zip_code_prefix	int64	first 5 digits of zip code	represent first five zip code of geolocation city and state
geolocation_lat	float64	Geolocation latitude	represent the angular distance of a place north or south of the earth's equator, or of a celestial object north or south of the celestial equator, expressed in degrees.
geolocation_lng	float64	Geolocation longitude	the angular distance of a place east or west of the meridian at Greenwich, England, or west of the standard meridian of a celestial object, expressed in degrees.
geolocation_city	object	city name	represent of geolocation city name
geolocation_state	object	state	represent 2 letter of state where the geolocation is
order_id	object	order unique identifier	An unique id where every transaction have order id

order_item_id	int64	sequential number identifying number of items included in the same order.	A serial number for an unique identifier assigned sequentially to an item, to uniquely identify it
product_id	object	product unique identifier	A serial number for a unique identifier assigned to an product, to uniquely identify it
seller_id	object	seller unique identifier	A serial number for a unique identifier assigned to an seller, to uniquely identify it
shipping_limit_date	object	Shows the seller shipping limit date for handling the order over to the logistic partner.	the estimated delivery date is when a customer can reasonably expect the order to arrive.
price	float64	item price	the retail price of each item of consumer goods.
freight_value	float64	item freight value item (if an order has more than one item the freight value is splitted between)	a price at which a certain cargo is delivered from one point to another
payment_sequential	int64	a customer may pay an order with more than one payment method. If he does so, a sequence will be	a sequential pay structure, principal payments are typically allocated to the securities in order of seniority.
payment_type	object	method of payment chosen by the customer.	A various of type payment for customer to pay a transaction
payment_installments	int64	number of installments chosen by the customer.	Instalment payments refer to a customer paying a bill in small portions throughout a fixed period of time
payment_value	float64	transaction value.	A total of amount money customer need to pay to successful transaction

III. Tiền xử lý

Dataset của Olist gồm 9 bảng nên ta phải nối các bảng lại.

```
df = orders.merge(customers, on = 'customer_id').merge(items, on = 'order_id')\
                                                 .merge(payments, on = 'order_id')\
                                                 .merge(reviews, on = 'order_id')\
                                                 .merge(products, on = 'product_id')\
                                                 .merge(trans, on = 'product_category_name')\
                                                 .merge(sellers, on = 'seller_id')
```

Sau đó, xoá cột tên product ngôn ngữ Portuguese thay bằng cột tên tiếng anh

```
df = df.drop('product_category_name', axis =1)
df.rename(columns = {"product_category_name_english ": "product_category_name"}, inplace = True)
```

Sau khi nối tất cả các bảng lại, chưa qua xử lí thì bộ dữ liệu gồm (115.609,39)

```
df.shape
✓ 0.1s
```

```
(115609, 39)
```

Tiền xử lý (thay những review_comment_title và message trống bằng no_tilte, no_comment, xoá null, duplicated). Sau khi tiền xử lý cơ bản ta thu được bộ dữ liệu với 94471 rows × 39 columns.

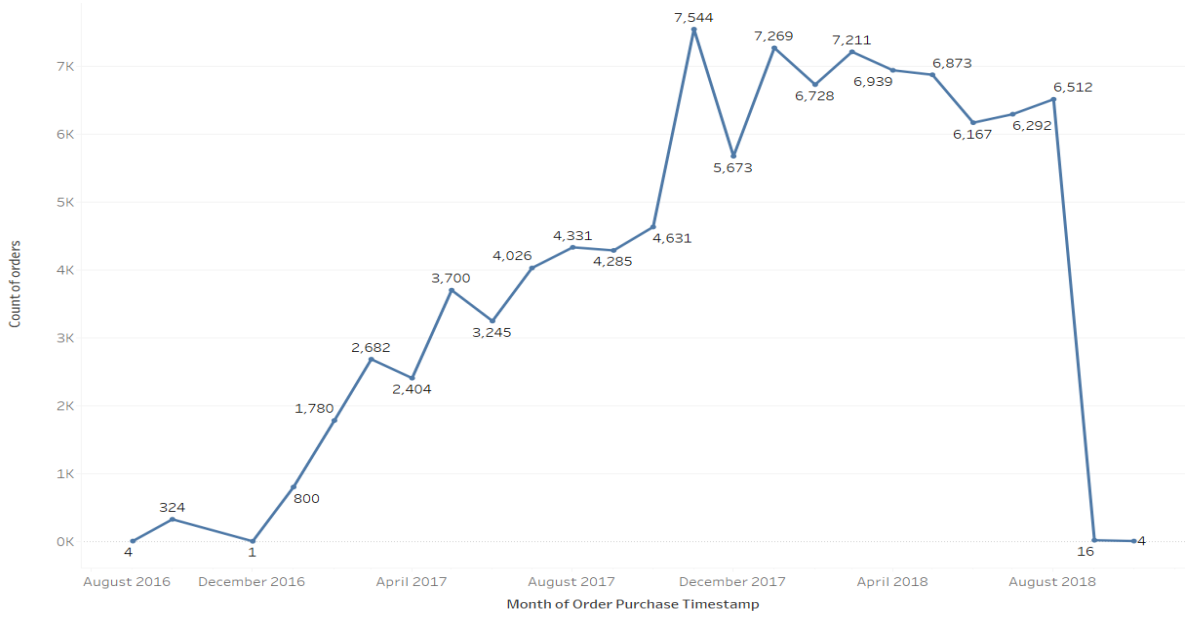
	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	10/2/2017 10:56	10/2/2017 11:07	10/4/2017 19:55
3	128e10d95713541c87cd1a2e48201934	a20e8105f23924cd00833fd87daa0831	delivered	8/15/2017 18:29	8/15/2017 20:05	8/17/2017 15:28
4	0e7e841ddf8f8f2de2bad69267ecfbcf	26c7ac168e1433912a51b924fbd34d34	delivered	8/2/2017 18:24	8/2/2017 18:43	8/4/2017 17:35
5	bfc39df4f36c3693ff3b63fcbea9e90a	53904ddbea91e1e92b2b3f1d09a7af86	delivered	10/23/2017 23:26	10/25/2017 2:14	10/27/2017 16:48
6	5f49f31e537f8f1a496454b48edbe34d	a7260a6ccba78544ccfaf43f920b7240	delivered	8/24/2017 11:31	8/24/2017 11:45	8/25/2017 14:17
...	...	...	...	...	...	...
115603	8dcb7601ceb0b144a5fdd0055b91ba28	6b9eb9660bed562d1c735d3fba0cfd60	delivered	6/19/2017 17:11	6/19/2017 17:25	7/19/2017 16:13
115604	edcf1e1eeb52381be9388c90152be52d	ce2172509c4149d65212484eb761bc37	delivered	8/21/2018 11:29	8/21/2018 11:50	8/21/2018 15:37
115605	2c12150c742ae2fa48bc703964c16c5f	ab0cf72dfe0538a63a57d6905ccb7b57	delivered	7/28/2018 17:55	7/29/2018 18:30	7/30/2018 14:54
115606	9f7ba492c273ffac8f90a623b8eadd3b	c61336d6c835632048f973d2f2db598d	delivered	8/23/2018 12:31	8/24/2018 3:24	8/24/2018 15:14
115607	2ef4a11b6e24fdfbb43b92cb5f95edff	ee1cfdc92e449920e25d3ca4ab4da4f6	delivered	7/23/2018 18:35	7/24/2018 10:31	7/24/2018 14:37

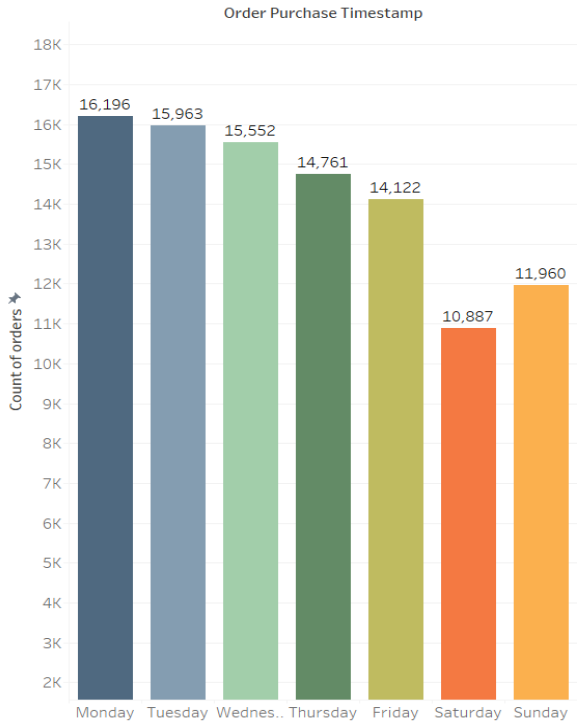
Hình minh hoạ Olist store dataset

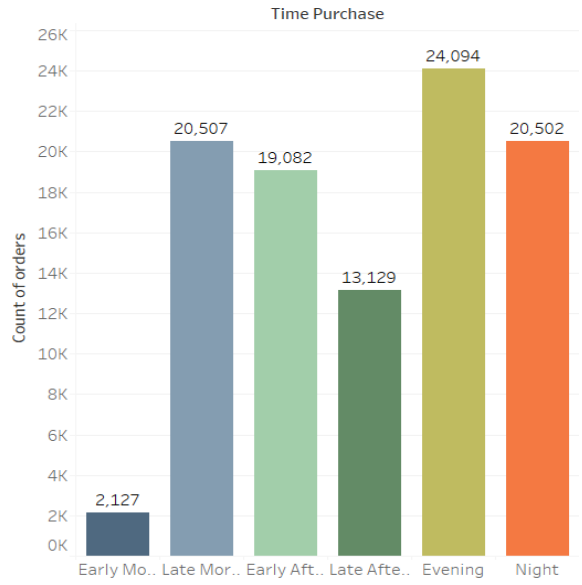
IV. Phân tích

1. Phân tích EDA: Tổng quan tình hình KH

- **Lượng đơn đặt hàng theo thời gian của Olist thay đổi như thế nào?**







Qua các biểu đồ trên, ta rút ra được:

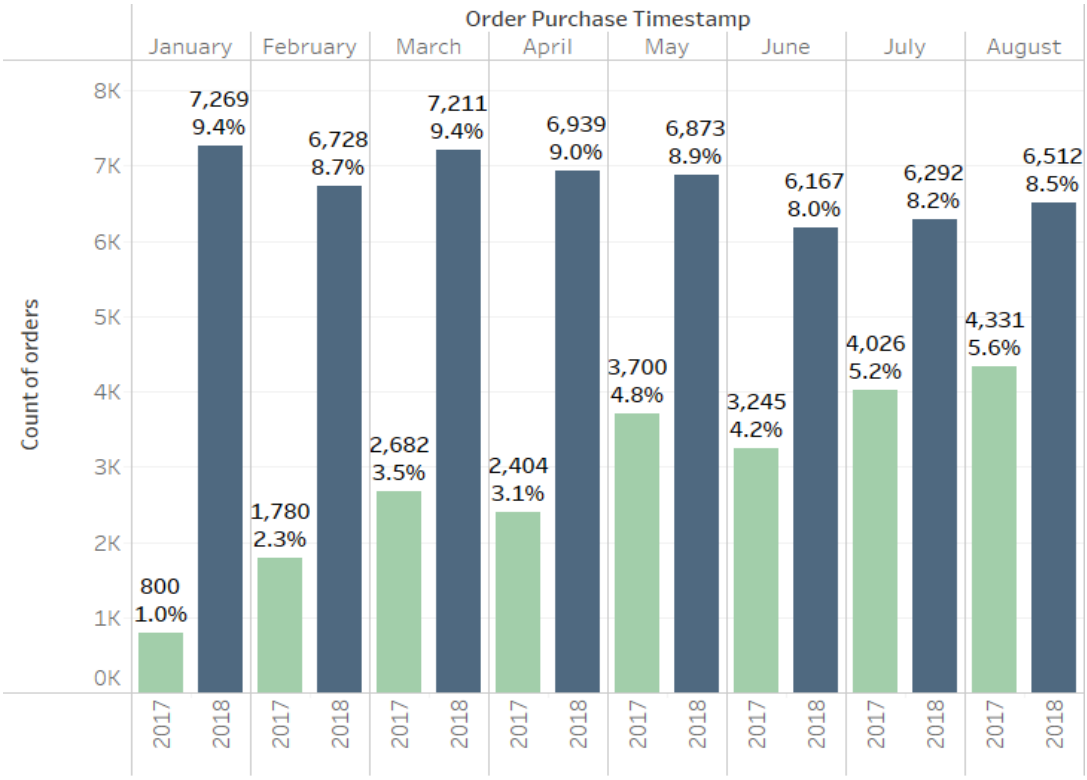
+ Olist có xu hướng phát triển theo thời gian. (dữ liệu ở phía cuối biểu đồ đường là outlier - dữ liệu nhiễu). Có thể thấy rõ ràng rằng khách hàng có xu hướng mua hàng trực tuyến nhiều hơn trước đây.
+ Đầu tuần là những ngày ưa thích của khách hàng Brazil và họ có xu hướng mua nhiều hơn vào buổi tối.

Qua quan sát có thể thấy đơn hàng sụt giảm lớn vào cuối tháng 8 năm 2018, đây là một hiện tượng lạ - có thể là outlier. Do đó, để tìm hiểu chi tiết, ta tiến hành so sánh dữ liệu các đơn đặt hàng từ tháng 1 - 8 năm 2017 và 2018

SL Order 2017-18 (tháng 1-8)

2017	2018
22,968	53,991

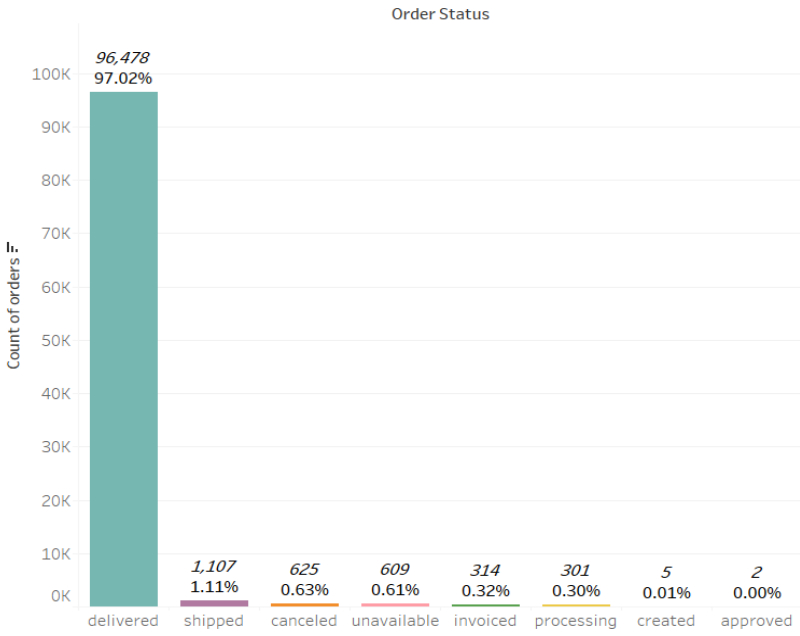
+ Số lượng đơn đặt hàng ghi nhận được năm 2017 (từ tháng 1 đến tháng 8) là gần 23 nghìn đơn, cũng với mốc thời gian trên thì 2018 thống kê được gần 54 nghìn đơn (*gấp hơn 2 lần* so với năm trước).



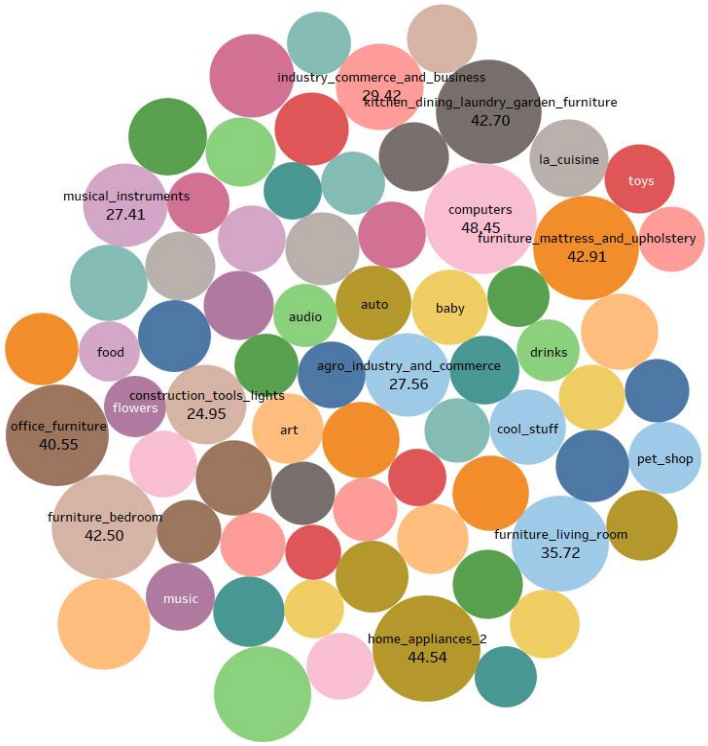
Có thể thấy, số lượng các đơn đặt hàng tăng theo thời gian. → Thực tế nguyên nhân của việc sụt giảm đơn vào cuối tháng 8 2018 vì dữ liệu Olist cung cấp dừng ở giai đoạn này.

- **Order Status**

Status Delivered - đã giao có số lượng đơn cao nhất, những tình trạng - status khác chỉ chiếm 3%.



- **Chi phí vận chuyển:** chi phí là một trong những mối quan tâm hàng đầu của KH vì nó ảnh hưởng trực tiếp đến giá cả mà họ phải chi trả.

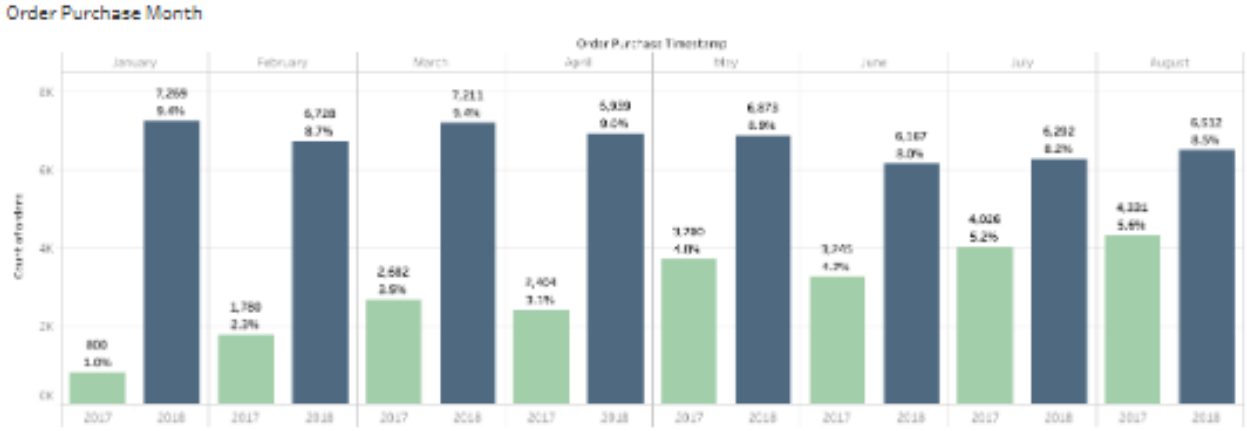


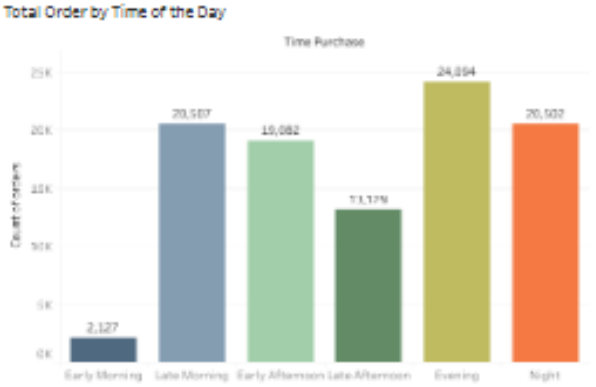
Chi phí vận chuyển phụ thuộc vào hình thức hàng hoá, phương thức vận tải, trọng lượng hàng hoá và khoảng các đến địa điểm giao hàng. Từ Packed Bubbles trên, ta có thể thấy, ***computers, home_appliances_2, furniture***,.. là nhóm những category có giá vận chuyển cao top đầu. Như chúng ta đã biết, trọng lượng là một trong những tiêu chí có ảnh hưởng lớn đến giá vận chuyển, do đó ***furniture, home_appliances*** có giá cao là điều hiển nhiên.

Để dõi thêm các thông tin khác về Olist, truy cập link Dashboard: https://public.tableau.com/views/Olist_Tableau/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link

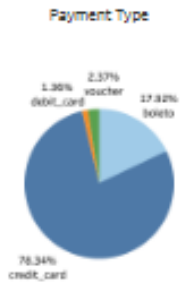
+ Payment Type
+ Review Score
+ Top product/ customer/ state
+ Forecast Sales

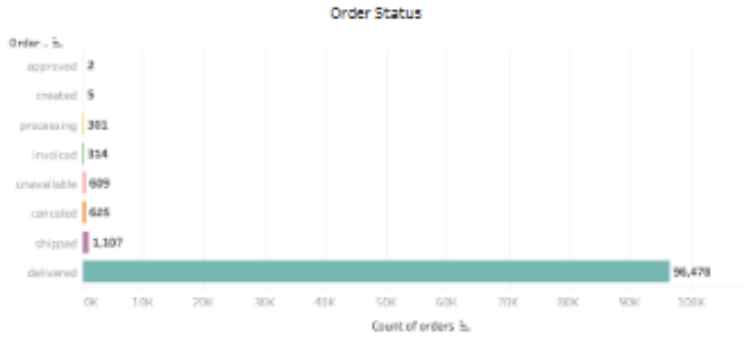
Olist_Tableau by Ngoc Xuan
Olist Store
Avg. Freight Value
19.99
Number of Orders
99,441
Avg. Payment
154.1
Select Year
(All)
Payment by Seller State
Sales by month
Order Delivered Customer Date
Payment Value
Year
2016
2017
2018
Seller State
Null
AC
AM
BA
CE
DF
ES
GO
MA
MG
MS
MT
PA
PB
PE
PI
PR
RJ
January
February
March
April
May
June
July
August
September
October
November
December


Order Purchase Month
Order Purchase Timestamp
January
February
March
April
May
June
July
August
Count of orders
2017
2018

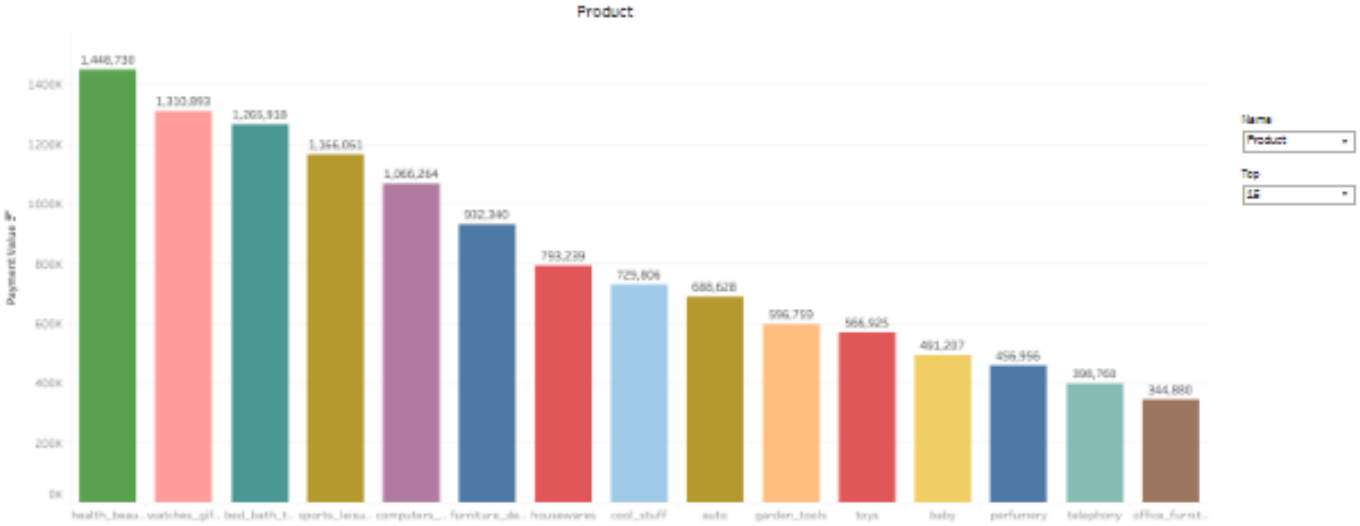

Total Order by Time of the Day
Time Purchase
Count of orders
Early Morning
Late Morning
Early Afternoon
Late Afternoon
Evening
Night

Order Weekly
Order Purchase Timestamp
Count of orders
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday


Payment Type
1.36%
debit_card
2.37%
voucher
17.92%
boleto
78.34%
credit_card


Order Status
approved 2
created 5
processing 301
invoiced 314
unavailable 609
canceled 625
shipped 1,107
delivered 96,478
0K 10K 20K 30K 40K 50K 60K 70K 80K 90K 100K
Count of orders

Review Score
Null 5,494
0 1
1 11,424
2 3,151
3 8,179
4 19,142
5 57,328
0K 5K 10K 15K 20K 25K 30K 35K 40K 45K 50K 55K 60K
Count of order_reviews


Product
1,448,730
1,330,893
1,265,918
1,166,061
1,066,264
932,340
793,239
725,806
688,628
596,759
564,925
491,207
456,956
398,768
344,880
Payment Value
1400K
1200K
1000K
800K
600K
400K
200K
0K
health_beau.. watches_gif.. bed_bath_t.. sports_leisu.. computers_.. furniture_de.. housewares cool_stuff auto garden_tools toys baby perfumery telephony office_furnit..
Name
Product
Top
15

Forecast
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
Count of customers
September 2016 January 2017 May 2017 September 2017 January 2018 May 2018 September 2018 January 2019 May 2019
Month of Order Delivered Customer Date
Forecast Indicator
Actual
Estimate

2. Customer Sentiment (count word)

Định nghĩa: Là cách khai thác nội dung theo từng ngữ cảnh được trích xuất từ các thông tin thu thập từ khách hàng, đồng thời giúp doanh nghiệp hiểu được mức độ cảm xúc của khách hàng đối với thương hiệu, sản phẩm và dịch vụ của họ khi nắm bắt được các feedbacks từ khách hàng. Đó là một dạng phân tích văn bản sử dụng quá trình xử lý ngôn ngữ tự nhiên (NLP) và máy học. Phân tích tình cảm còn được gọi là "khai thác ý kiến" hoặc "trí tuệ nhân tạo cảm xúc".

Sentiment Scoring:

Chìa khóa của phân tích cảm xúc là phân loại phân cực. Phân loại phân cực đề cập đến cảm xúc tổng thể bao quát cả một đoạn text, một câu hoặc thậm chí là vài từ. Kiểu phân loại này được thể hiện là rating dưới dạng số gọi là sentiment scoring. Ví dụ, điểm số này là một số từ -100 đến 100 trong đó số 0 thể hiện cảm xúc trung lập, trên 0 là biểu hiện tích cực và số âm biểu thị cảm xúc tiêu cực dành cho sản phẩm đó. Điểm số này được tính cho toàn bộ văn bản hoặc các câu đơn lẻ. Ngoài ra chúng ta có thể sử dụng Fine-grained Sentiment Analysis để chi tiết hóa mức độ cảm xúc của khách hàng (Rất tiêu cực, tiêu cực, trung tính, tích cực và rất tích cực), thậm chí có thể chi tiết hơn nữa, ví dụ trong sự tích cực có thể phân thành các cảm xúc như hạnh phúc, phấn khích, ấn tượng,... nhưng trong bài chúng ta chỉ sử dụng Sentiment Scoring là đủ vì chúng ta chỉ cần biết cảm xúc cố định của khách hàng là như thế nào để đưa ra các giải pháp đối với mặt hàng mà khách hàng đánh giá.

Thực hiện phân tích trên JupyterLab: gồm

- Đầu tiên chúng ta tiền xử lý trước khi bắt đầu xây dựng mô hình, đối với bước này chỉ dùng dataset olist_order_reviews_dataset.csv và olist_products_dataset.csv.

Với dataset review: vì dữ liệu dạng text đều viết bằng tiếng Portuguese nên chúng ta sẽ chuyển đổi toàn bộ cột sang tiếng anh để dễ phân tích và trực quan hơn.

+ Mình dùng thư viện googletrans để import vào sau đó chuyển frame thành dạng string để dịch sang tiếng anh toàn bộ điểm dữ liệu trong data:

```
'0 NAN 1 NAN 2 NAN 3 I received well before stipulated period.4 Congratulations Lannister Stores I loved buying i ... ... 99
219 NAN 99220 NAN 99221 Excellent backpack, super fast delivery.Super ... 99222 nan 99223 Product arrived already returned,
po ... name: review_comment_message, length: 99224, dtype: object'
```

Tiếp tục, chuyển nó lại thành dạng dataframe rồi nối với data

product_id	review_eng
1e9e8ef04dbcff4541ed26657ea517e5	Super adooorei eyeliner it's very black and I ...
3aa071139cb16b67ca9e5dea641aaa2f	Excellent pointing! In 3/2 a perfect tip!
96bd76ec8810374ed1b65e291975717f	NaN

olist_products_dataset.csv để phân tích comment trên từng sản phẩm.

+ Tiếp theo mình sử dụng Regex để xóa emoji trên từng review để thuận tiện hơn cho việc tách từ

```
def remove_emoji(string):
    emoji_pattern = re.compile("["
                            u"\U0001F600-\U0001F64F"
                            u"\U0001F300-\U0001F5FF"
                            u"\U0001F680-\U0001F6FF"
                            u"\U0001F1E0-\U0001F1FF"
                            u"\U00002702-\U000027B0"
                            u"\U000024C2-\U0001F251"
                            "]+", flags=re.UNICODE)
    return emoji_pattern.sub(r'', string)
reviews.review_eng = reviews.review_eng.apply(lambda text: remove_emoji(str(text)))
```

+ Tiếp theo, mình sử dụng Word-based tokenization để tách các từ trong câu, sử dụng Stop words để loại bỏ các từ gây nhiễu và xóa các text bị NaN. Cuối cùng là dùng Regex để xóa các ký tự đặc biệt, kết quả:

	product_id	review_eng	review_score
21424	7501e40e8bc7a7db7c2f4b005e1dddcb	forwards received great condition	4
11700	c5edc6d6a54cbcc12367409391fdfc50	good	5
9060	656e0eca68dcecf6a31b8ececfabe3e8	long wait	1

+ Sử dụng Sentiment Scoring dựa trên thư viện Afinn của Python để tiến hành phân ra đánh giá mỗi sản phẩm bao gồm: negative, neutral và possitive, cụ thể nếu mỗi review có điểm < 0 thì là negative, điểm = 0 là neutral và còn lại là possive.

	product_id	reviews	score	sentiment
11089	293ee8ac0a45f559cdeba37b938fdb66	product quality could little high side overall...	5.0	positive
10173	7c1bd920dbdf22470b68bde975dd3ccf	congratulations	2.0	positive
11525	736f1b87428f9cfe5f5184c4ac0fbe05	good product delivered within postal strike	2.0	positive
21916	6e7b435ac3c84fc5348dba974e184322	good	3.0	positive
21387	fec85e6c0eab9b4bcddf9ccd917c80bc	ok product delivery	0.0	neutral
24992	f9cce3e19235fa7d11942f69caaa22c7	came crumpled small cap battery complained wou...	-6.0	negative
207	a3c95dc653d9a279b160bc0179385ed1	good product	3.0	positive

Qua bảng trên ta thấy được mô hình Afinn dự báo rất tốt dựa trên từ ngữ feedback từ khách hàng, chẳng hạn như sản phẩm có id 736f1b87428f9cfe5f5184c4ac0fbe05 với review là good product delivered within postal strike có chỉ số Scoring Sentiment tương đối ổn với cảm xúc được đánh giá là possitive còn với mặt hàng f9cce3e19235fa7d11942f69caaa22c7 thì reviews lại tệ hơn "came crumpled small cap battery complained…" có chỉ số khá thấp là -6.0 được đánh giá là tiêu cực.

+ Việc tiếp theo mình sẽ nhóm các sản phẩm có review theo từng loại vào mỗi nhóm xếp từ positive -> neutral -> negative. Dưới đây là các sản phẩm được review tích cực từ khách hàng:

product_id	score	sentiment
37664a6100bf523c545b7040be69ad00	17.0	positive
f71954485428cf87945953ea0a0a6229	16.0	positive
f82a4b08cf7b2bf375fb77e519231f9a	16.0	positive
c696cd2db7870be08c37822f7c0ef831	15.0	positive
92857a2b590adeae8d5992511dc197e9	15.0	positive
1574ed2c73e4465c572f68dd77528203	14.0	positive
12e479f1c2c0417147a3e36302485e30	14.0	positive
4563095e06df1fa67de2eade86f4f01b	14.0	positive

Visualize words của các sản phẩm

Qua Wordcloud, mình thấy được các từ ngữ biểu thị sự hài lòng như super, recommend, perfect, great, beautiful,... thông qua các sản phẩm được dự đoán là Olist đem lại sự hài lòng cho khách hàng thông qua phân tích cảm xúc. Bên cạnh đó, wordcloud còn cho biết khách hàng của Olist đặc biệt chú trọng đến *delivery* - thời gian, địa điểm giao hàng. Với những mặt hàng này, doanh nghiệp cần phải tập trung tối đa hóa lợi ích của nó dành cho khách hàng và cải thiện, nâng cao chất lượng các lĩnh vực khách hàng quan tâm như: hình thức mẫu mã, các vấn đề liên quan đến giao hàng, chất lượng cũng như giá cả. Cùng với việc tăng cường số lượng các sản phẩm đó.

Về vấn đề giao hàng mình sẽ dùng vài kĩ thuật chi tiết hơn để đánh giá thái độ khách hàng đối với yếu tố này:

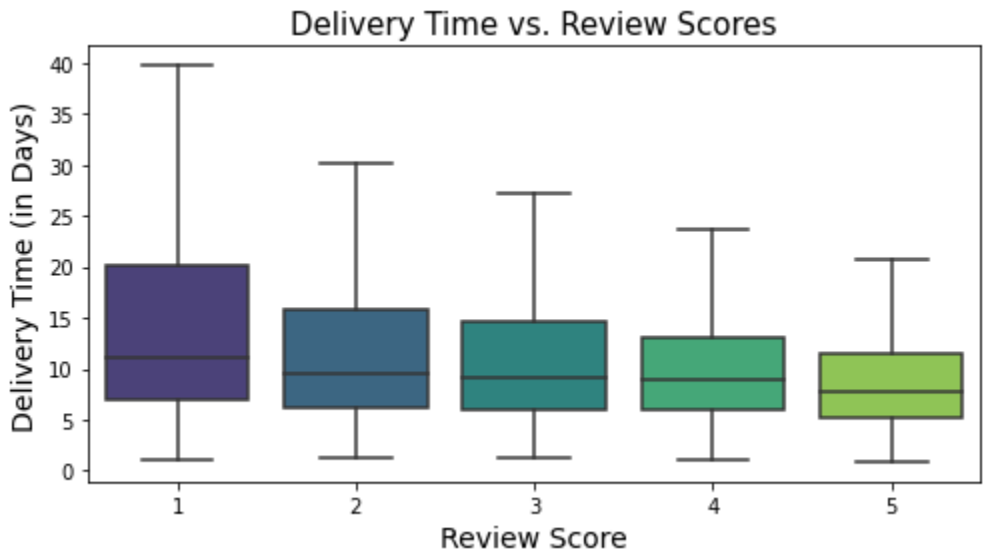
- Dựa vào Review score được đánh giá trên mỗi sản phẩm mà khách hàng nhận được.

- Dựa vào vị trí địa lý, số lượng hàng hóa của mỗi khu vực.

Tính thời gian giao hàng trên mỗi đơn hàng:

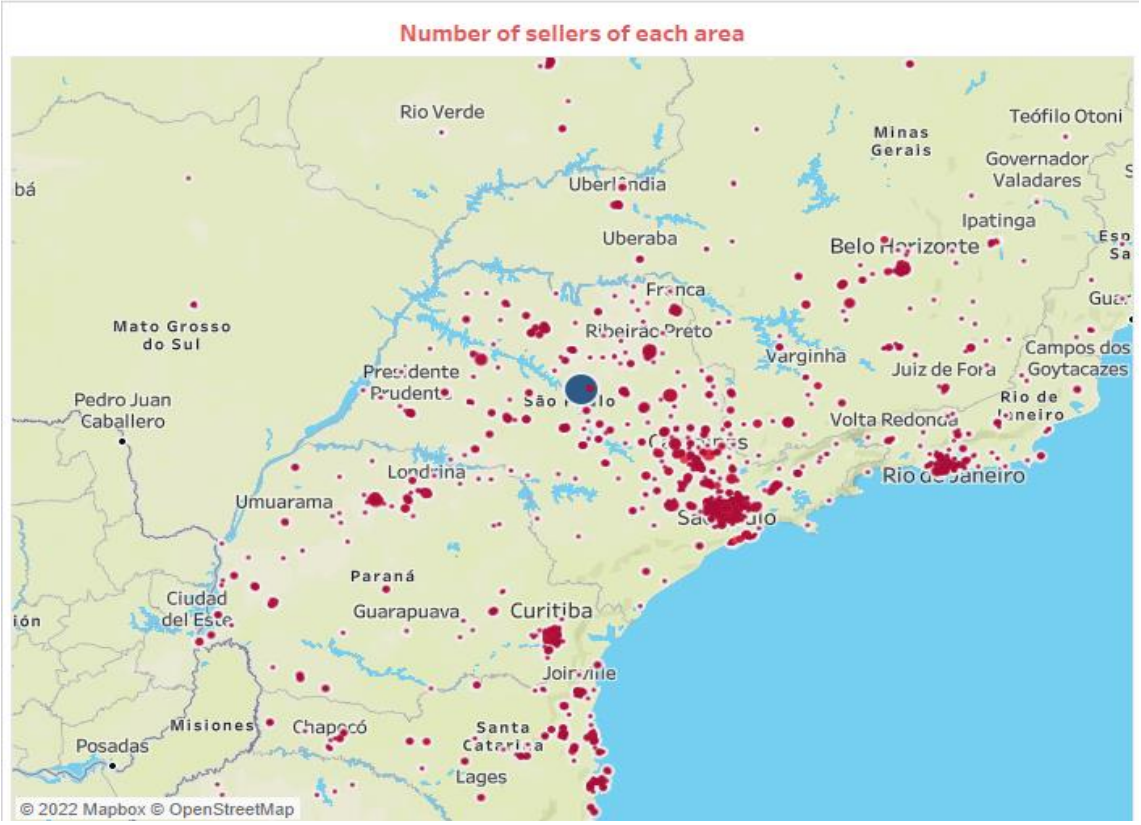
```
deliveryTime = (df["order_delivered_customer_date"]-df["order_purchase_timestamp"])
deliveryTime_Seconds = deliveryTime.apply(lambda x: x.total_seconds())
df['deliveryTime_Days'] = round(deliveryTime_Seconds/(24*3600), 2)
```

Tiếp theo mình dùng Box plot để trực quan mối quan hệ giữa thời gian giao hàng và điểm đánh giá của khách hàng trên tất cả các đơn hàng:



Với biểu đồ trên, các đơn hàng được giao từ 15 đến 20 ngày đều nhận đánh giá khá thấp từ khách hàng và ngược lại, các đơn hàng có thời gian giao hàng nhanh (dao động từ 7 đến 12 ngày) được phản hồi bằng những đánh giá khá tích cực (từ 4 đến 5 điểm đánh giá). Và chi tiết hơn, chúng ta sẽ tìm nguyên nhân dẫn đến vấn đề giao hàng chậm trễ và đề xuất các giải pháp giúp cho doanh nghiệp phần nào giải quyết được vấn đề này.

➔ Dùng Tableau để trực quan bản đồ địa lý nơi cung cấp hàng hóa cho thị trường (các kho hàng)



Số lượng kho hàng tại các khu vực

Chỉ số đánh giá Sentiment Scoring theo từng khu vực

Theo bản đồ thứ 2 so với bản đồ 1 ta thấy được khu vực *Sao Paulo* là có chỉ số positive cao nhất (màu xanh), xung quanh khu vực này cũng khá cao. Dễ dàng nhận thấy một yếu tố ảnh hưởng khá lớn đến chỉ số hài lòng của khách hàng đó là vấn đề shipping (dựa vào word count negative).

Đối với bản đồ kho hàng (seller) ta thấy *Sao Paulo* và *khu vực xung quanh Sao Paulo* có số kho hàng local rất nhiều, lớn, do đó, các vấn đề liên quan việc ship hàng hoạt động hiệu quả, đi đôi với việc số lượng đơn hàng, số lượng sản phẩm được đặt rất cao →chỉ số positive rất cao (đứng đầu khu vực) và đem lại doanh thu lớn nhất khu vực. Còn đối với các khu vực như *Mato Grosso do Sul, Amapa,...* có "rất ít kho hàng" và nằm khá "xa" khu vực tập trung khách hàng và khu vực nhiều đơn hàng cộng thêm địa hình hiểm trở (dựa trên biểu đồ thứ 2) cho nên lương đơn đặt hàng ít và chỉ số đánh giá negative rất cao (theo chiều âm)

Đề xuất: Có thể bổ sung/ xây thêm kho hàng ở các khu vực này và liên kết với các nhà bán lẻ ở các khu vực đó để giải quyết vấn đề ship hàng → tăng được sự hài lòng cho khách hàng, tăng lượng đơn đặt hàng nhằm tăng doanh thu và mở rộng vị thế trên tất cả mọi khu vực.

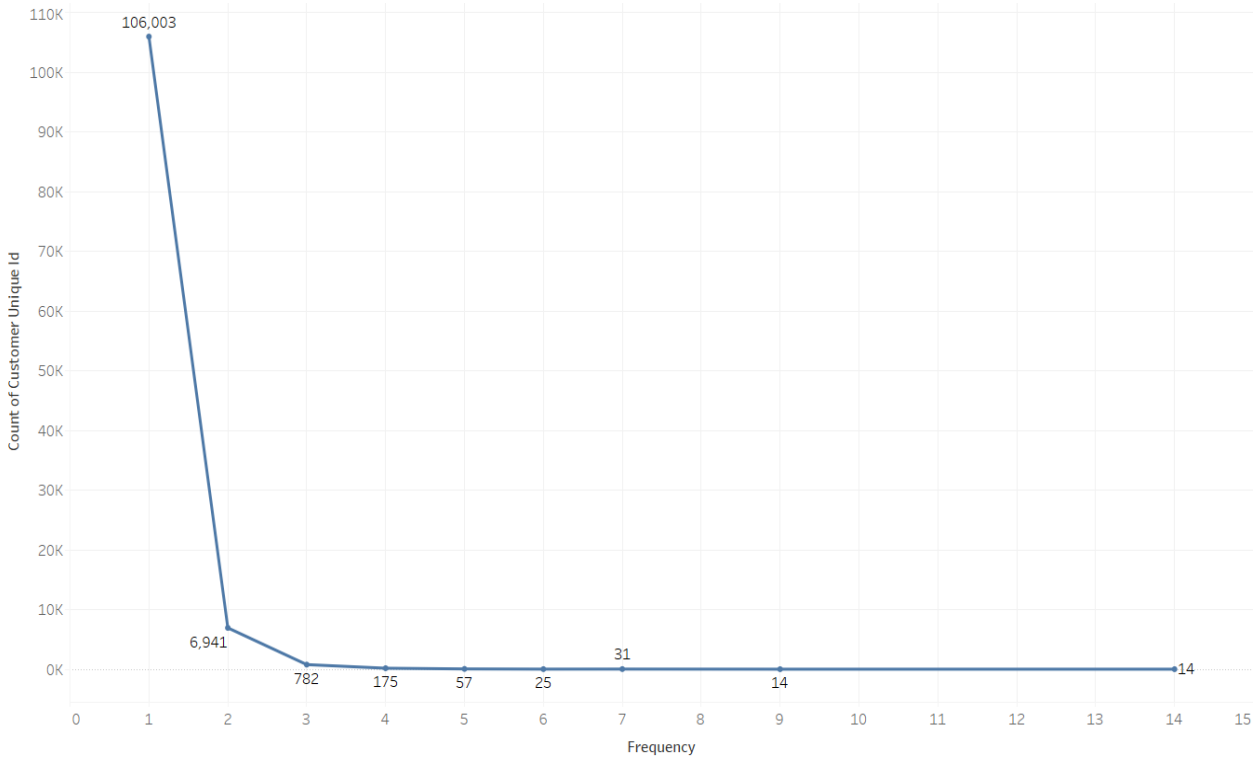
3. Customer Segmentation

Phân nhóm khách hàng là hành động tìm kiếm và xác định những đặc điểm chung của các khách hàng, đây là một trong những chiến lược chăm sóc khách hàng hiệu quả. Bởi giúp doanh nghiệp nắm bắt chính xác nhất những đặc điểm chung trong dữ liệu của tệp khách hàng.

Bên cạnh đó, thực hiện phân nhóm giúp thống kê các đặc tính tương đồng giữa những khách hàng trước và sau khi mua sản phẩm, dịch vụ tại doanh nghiệp. Từ đó, khách hàng sẽ được phân chia thành từng nhóm nhỏ, tạo sự khác biệt trong nhu cầu khách hàng và giúp mang lại những trải nghiệm hài lòng nhất cho họ.

Tần suất mua hàng của khách hàng như thế nào ?

Qua biểu đồ bên có thể thấy, hầu hết khách hàng chỉ đến mua 1 lần (>100.000), lượng khách quay lại mua lần 2,3 rất ít.



Tần suất mua hàng của khách hàng Olist

Sử dụng RFM và Kmeans để phân loại khách hàng.

3.1 RFM:

	customer_unique_id	Recency	Frequency	Monetary
4310	7564d75e79ac40aac2a1a0e1e9fd9a2a	313	1	50.00
5420	937a668a37820388e591a0bfc2392d08	310	1	76.95
8030	dc460c8c5a048d071e9fe80391c1e4da	292	1	133.76
370	09bf8d9eeaf0516be20397d0a90ddbd2	271	1	116.00
5601	97e1b81bc233b224f3355ad9bc667d46	260	1	30.00
5492	954787ed2823ee776c15d070b287c4d9	239	1	36.69
4802	823baafc843bd105b854214889d74649	236	1	180.50

3.2. Dùng Kmeans để phân cụm khách hàng:

a. Chuẩn hóa Recency, Frequency, Monetary theo StandardScaler

customer_unique_id	Recency	Frequency	Monetary
0000366f3b9a7992bf8c76cfdf3221e2	0.896227	-0.100373	-0.146364
000ec5bff359e1c0ad76a81a45cb598f	-1.553025	-0.100373	-0.610280
00172711b30d52eea8b313a7f2cced02	-0.982325	-0.100373	-0.226955

b. Lọc các giá trị bất thường.

Recency:

	customer_unique_id	Recency	Frequency	Monetary
67181	b7d76e111c89f7ebf14761390f0f7d17	3.175374	-0.160426	0.093539
26477	4854e9b3feff728c13ee5fc7d1547e92	3.168845	-0.160426	-0.219035
5373	0eb1ee9dba87f5b36b4613a65074337c	2.992564	-0.160426	-0.164754
17384	2f64e403852e6893ae37485d5fcacdaf	2.986035	-0.160426	-0.275992
42275	7390ed59fa1febbfda31a80b4318c8cb	2.986035	-0.160426	-0.093135

Không có giá trị quá khác biệt nên không xoá.

Frequency:

	customer_unique_id	Recency	Frequency	Monetary
51506	8d50f5eadf50201ccdcedfb9e2ac8455	-1.492809	67.071700	0.960781
22823	3e43e6105506432c953e165fb2acf44c	-0.356775	38.257931	2.771350
10077	1b6c7548a2a1f9037c1fd3ddfed95f33	-0.271899	28.653342	1.857633
73978	ca77025e7201e3b30c44b472ff346268	-0.970495	28.653342	3.029231
36755	6469f99c1f9dfae7733b25662e7f1782	-1.146776	28.653342	1.202954

Không có giá trị quá khác biệt nên không xoá.

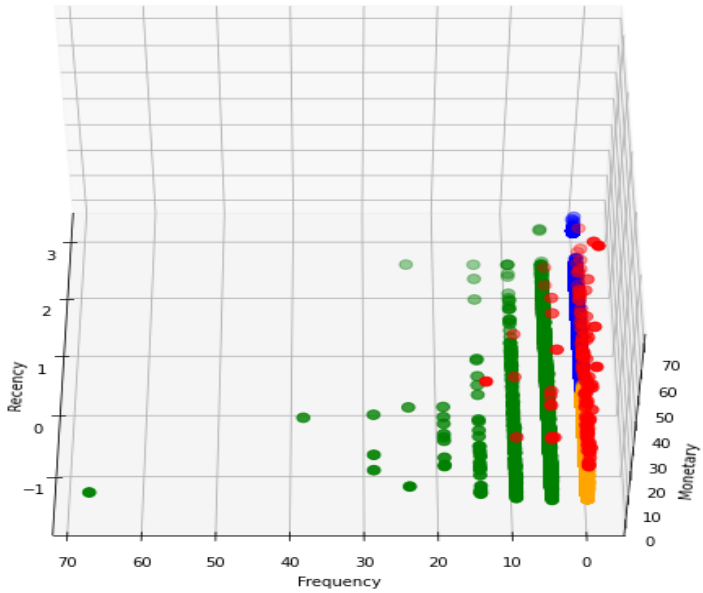
Monetary:

	customer_unique_id	Recency	Frequency	Monetary
3724	0a0a92112bd4c708ca5fde585afaa872	0.629092	-0.160426	172.753663
71634	c402f431464c72e27330a67f7b94d4fb	-0.324131	-0.160426	69.410099
23445	4007669dec559734d6f53e029e360987	0.263473	-0.160426	57.441149
87489	ef8d54b3797ea4db1d63f0ced6a906e9	-0.689751	-0.160426	47.460258
43207	763c8b1c9c68a0229c42c9fc6f662b93	-1.257767	-0.160426	45.739866

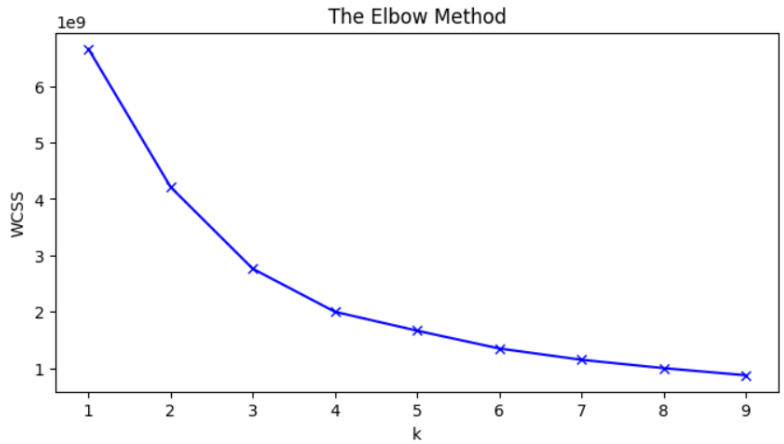
Có khách hàng số 3724 mua với số tiền rất lớn (không ảnh hưởng đến kinh doanh) nhưng cần lọc ra để dễ phân cụm và gán biến cho nó là Super. Điểm dữ liệu này này để sau khi ta dùng Kmeans sẽ bỏ lại vào mô hình chứ không nên xóa đi.

```
super = rfm.sort_values('Monetary', ascending = False).iloc[0:1]
```

Kết quả đã phân được các cụm khách hàng theo RFM bằng Kmeans:



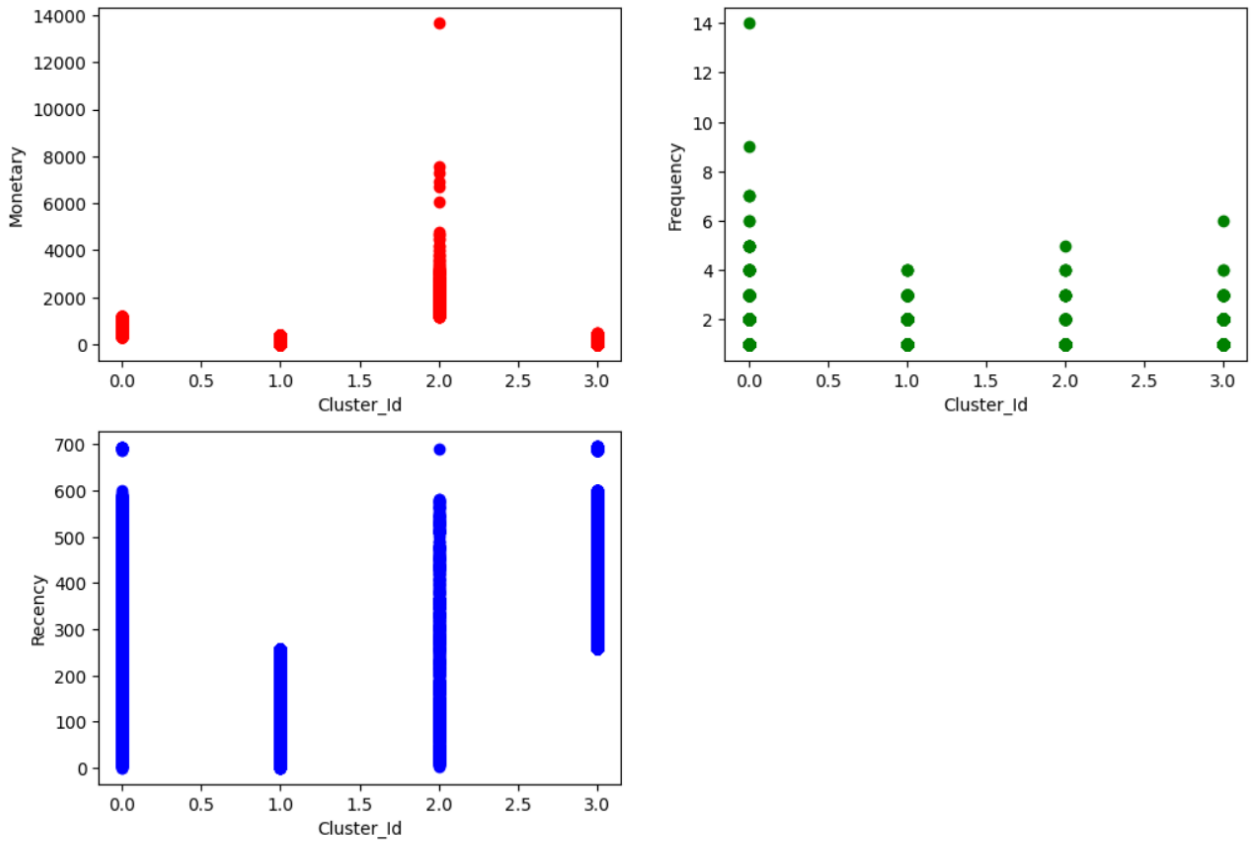
Dựa vào phương pháp Elbow Method và mong muốn, ta sẽ chia customer thành 4 cụm theo RFM (Recency - Frequency - Monetary) tính toán được.



	customer_unique_id	Recency	Frequency	Monetary	Cluster_Id	customer_segment
0	0000366f3b9a7992bf8c76cfdf3221e2	111	1	141.90	2	valuable
1	0000b849f77a49e4a4ce2b2a4ca5be3f	114	1	27.19	2	valuable
2	0000f46a3911fa3c0805444483337064	537	1	86.22	2	valuable
3	0000f6ccb0745a6a4b88665a16c9f078	321	1	43.62	2	valuable
4	0004aac84e0df4da2b147fca70cf8255	288	1	196.89	2	valuable
...	...	...	...	...	...	...
91460	fffcf5a5ff07b0908bd4e2dbc735a684	447	1	2067.42	3	passive
91461	fffea47cd6d3cc0a88bd621562a9d061	262	1	84.58	3	passive
91462	ffff371b4d645b6ecea244b27531430a	568	1	112.46	1	occasion
91463	ffff5962728ec6157033ef9805bacc48	119	1	133.69	1	occasion
91464	ffffd2657e2aad2907e67c3e9daecbeb	484	1	71.56	3	passive

- **Đặc điểm của mỗi cụm và đề xuất chiến lược phù hợp**

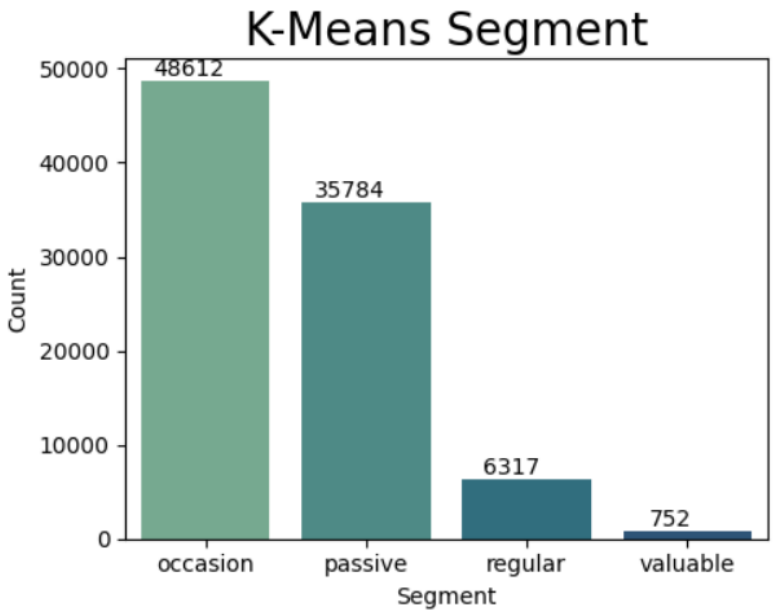
Sau đó, để xác định thuộc tính, những điểm đặc trưng riêng của mỗi cụm, ta vẽ biểu đồ để so sánh. Qua biểu đồ bên dưới, ta có thể nhận xét:



+ *Cụm 0:* Nổi bật với tần suất mua hàng thường xuyên - nhiều nhất so với 3 cụm còn lại → **Regular - những khách hàng mua thường xuyên**
+ Cụm 1: Phần lớn khách hàng mua gần đây với tần suất thấp (1 lần), thỉnh thoảng mua và số tiền không lớn → **Occasion- những khách hàng nhân dịp**
+ Cụm 2: Chi tiêu với số tiền lớn hơn nhiều so với các cụm còn lại → **Valuable - những khách hàng giá trị.**
+ Cụm 3: Số tiền khách hàng chi tiêu ít, tần suất mua hàng cũng rất thấp (1 lần) và hầu hết khách hàng đến cửa hàng từ > 200 đến 600 ngày trước. Nên ta đặt tên cho nhóm này là **Passive - những khách hàng thụ động**

- **Thống kê số lượng khách hàng của mỗi cụm**

Có thể thấy khách hàng chủ yếu của Olist Store đến từ nhóm occasion và passive với đặc điểm là tần suất mua hàng rất thấp, số tiền bỏ ra ít. Do đó, Olist cần phải đưa ra các chiến lược marketing chăm sóc khách hàng phù hợp, tăng độ nhận diện, gợi nhớ để thu hút khách hàng quay lại và chi tiêu nhiều hơn.



Bên cạnh đó, những khách hàng regular và valuable cũng có vai trò rất quan trọng.

Recommend:

- *Đối với Passive Customers:*

+ Tạo nhận thức, độ nhận diện tốt về các sản phẩm, quảng cáo nhiều hơn. Giúp khách hàng biết đến và có nhiều lựa chọn, mua thêm những mặt hàng khác.
+ Giảm giá thanh toán bằng credit card và boleto. Áp dụng với các sản phẩm họ thường mua (health beauty product, computer accessories, and houseware - bed bath table and sport leisure)
+ Cung cấp các voucher vận chuyển miễn phí cho hàng hoá vận chuyển trong bang và gần đó.
+ Tăng thêm các sự kiện giảm giá - ngày lễ như black friday

- *Đối với Regular Customers:*

+ Thường xuyên giới thiệu lợi ích khi trở thành khách hàng valuable-loyal Customer
+ Đưa ra đề xuất sản phẩm mới dựa vào lần mua trước
+ Voucher giảm tiền khi họ mua thêm sản phẩm từ Olist.
+ Tặng thẻ Bạc kèm theo các chương trình, mã giảm hấp dẫn.

- *Đối với Occasion:*
+ Khuyến mãi khi mua hàng lần thứ 2, điều này sẽ thúc đẩy khách hàng đi mua hàng mới, giúp tăng tần suất mua hàng.
+ Giảm giá thanh toán bằng credit card và boleto. Áp dụng với các sản phẩm họ thường mua (health beauty product, computer accessories, and houseware - bed bath table and sport leisure)

- *Đối với Valuable Customers:*
+ Voucher giảm tiền khi họ mua thêm sản phẩm từ Olist.
+ Chiết khấu freigh_value
+ Tặng thẻ khách hàng Kim cương (có các discount hấp dẫn hàng tuần, tháng)

Link tableau:

1. Tổng quan
2. Sentiments