olist

A LOOK AT CUSTOMER LIFETIME VALUE

Rachelle Perez

December 2019

WHO IS OLIST?

- **E-Commerce** business
- Small business merchants ("sellers") **sell their products** to customers through Olist and **ship them directly** to customer **using Olist logistics partners** ("carrier")
- olist.com

olist

PROBLEM: What **factors** affect 6-month Customer Lifetime Value (LTV)?

PROJECT: Build a **regression model** that explains correlations

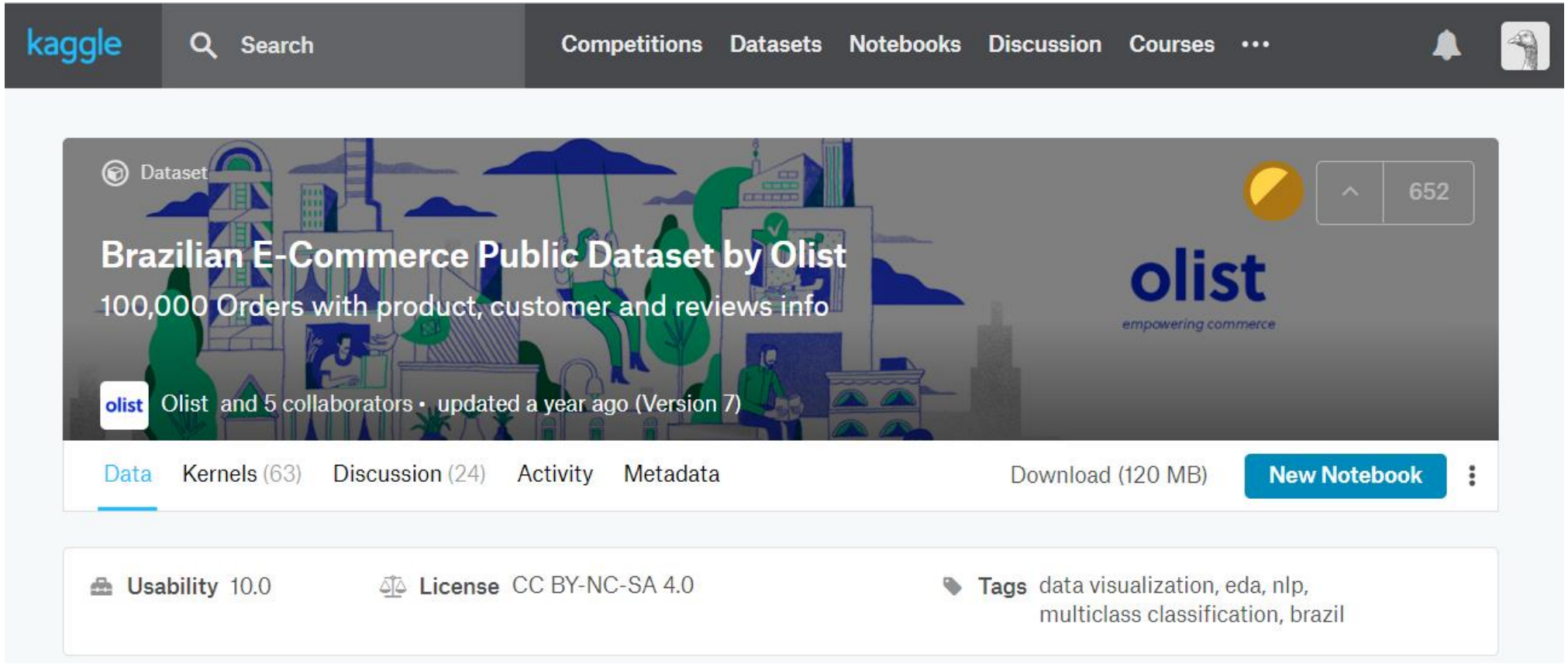
DATA PREPARATION

What data do we have?

What variables will be added to model?

What changes are needed for data to be ready for model?

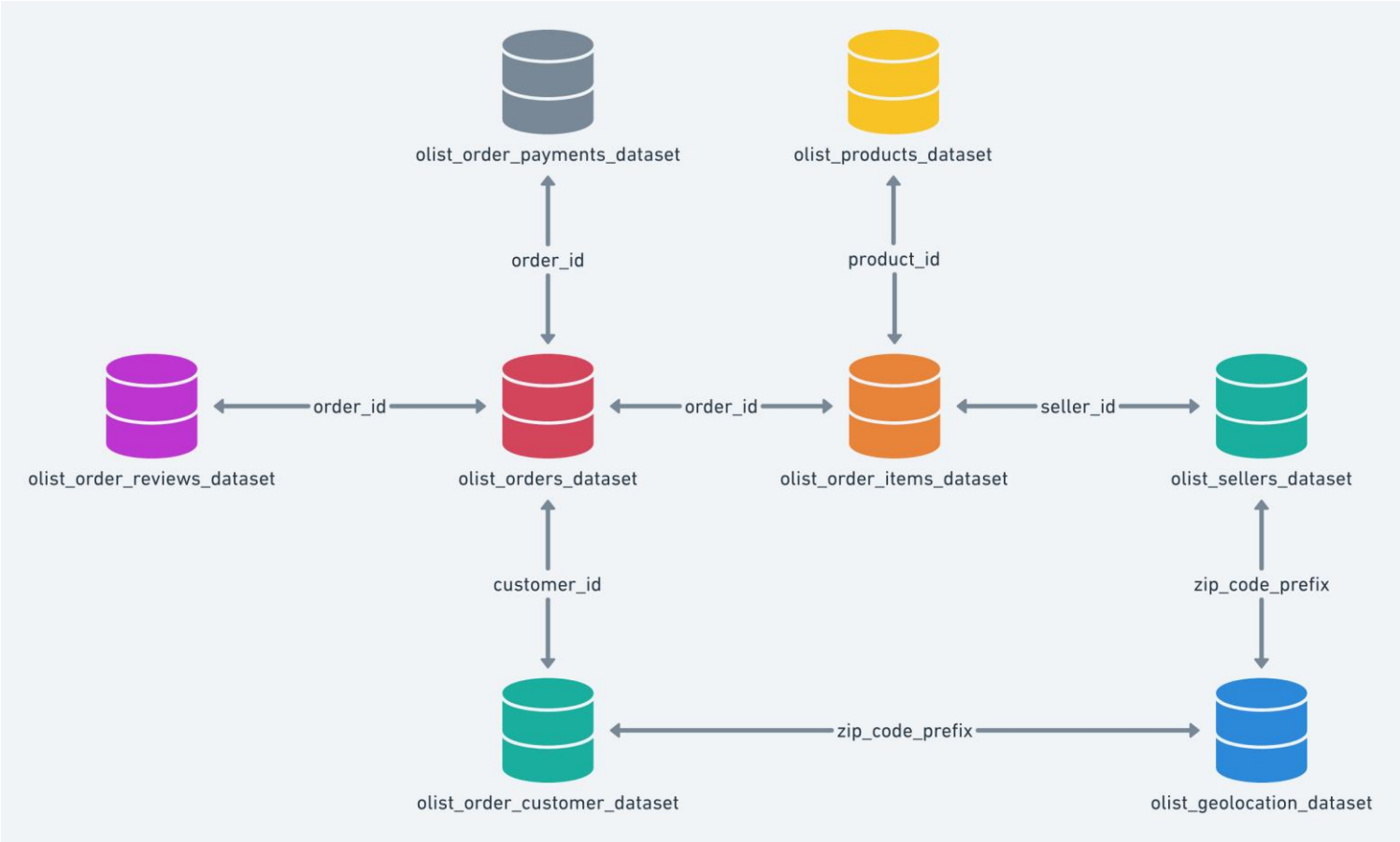
DATA ACQUISITION



https://www.kaggle.com/olistbr/brazilian-ecommerce

DATA PREVIEW

SCHEMA



INCLUDES

- Date Range: 9/4/2016 – 8/29/2018
- 96,096 Customers
- 99,441 Orders
- 32,951 Unique Products
- 3,095 Sellers

CHALLENGES

- Dependent variable (LTV) by customer and the only customer variables given are city, state, and zip. Aggregates must be created
- Aggregates are difficult as 1) Schema not linear 2) data is split for each **customer**, each **order** per customer, each **product** per order, and each **item** per product

DATA CLEANING

POSTGRESQL

Column	Type
customer_unique_id	character varying(50)
date_first_order	timestamp without time zone
year_first_order	double precision
month_first_order	double precision
total_orders_first_6_months	bigint
total_paid_first_6_months	double precision
order_count_unavailable	bigint
avg_payment_processing_time	interval
avg_seller_processing_time	interval
avg_transit_time	interval
avg_lead_time	interval
avg_item_count_per_order	numeric
avg_product_count_per_order	numeric
orders_shipped_late	bigint
avg_quantity_by_product	numeric
average_price_per_unit	double precision
avg_freight_cost_per_order	double precision
ordered_from_top_10_prod_category_bol	integer
ordered_from_seller_perfect_avg_review_bol	integer
avg_survey_lag	interval
avg_review_lag	interval
avg_review_score	numeric
order_count_boleto_voucher	bigint
order_count_card	bigint
avg_installments	numeric
avg_days_payment_processing_time	double precision
avg_days_seller_processing_time	double precision
avg_days_transit_time	double precision
avg_days_lead_time	double precision
avg_days_survey_lag	double precision
avg_daysreview_lag	double precision

- Create dataframe with aggregates
- Remove cancelled orders & orders not from customer's first 6 months

PYTHON

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 95557 entries, 0 to 95556
Data columns (total 31 columns):
customer_unique_id                        95557 non-null object
date_first_order                          95557 non-null object
year_first_order                          95557 non-null int64
month_first_order                         95557 non-null int64
total_orders_first_6_months               95557 non-null int64
total_paid_first_6_months                 95557 non-null float64
order_count_unavailable                   95557 non-null int64
avg_payment_processing_time               95538 non-null object
avg_seller_processing_time                94217 non-null object
avg_transit_time                          93306 non-null object
avg_lead_time                             93327 non-null object
avg_item_count_per_order                  94978 non-null float64
avg_product_count_per_order               94978 non-null float64
orders_shipped_late                       85918 non-null float64
avg_quantity_by_product                   94978 non-null float64
average_price_per_unit                    94978 non-null float64
avg_freight_cost_per_order                94978 non-null float64
ordered_from_top_10_prod_category_bol     95557 non-null int64
ordered_from_seller_perfect_avg_review_bol 95557 non-null int64
avg_survey_lag                            85569 non-null object
avg_review_lag                            88879 non-null object
avg_review_score                          95557 non-null float64
order_count_boleto_voucher                22800 non-null float64
order_count_card                          75113 non-null float64
avg_installments                          95557 non-null float64
avg_days_payment_processing_time          95538 non-null float64
avg_days_seller_processing_time           94217 non-null float64
avg_days_transit_time                     93306 non-null float64
avg_days_lead_time                        93327 non-null float64
avg_days_survey_lag                       85569 non-null float64
avg_daysreview_lag                        88879 non-null float64
dtypes: float64(17), int64(6), object(8)
memory usage: 22.6+ MB
```

WRONG TYPE

COUNTS

NULLS

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 95557 entries, 0 to 95556
Data columns (total 37 columns):
customer_unique_id                        95557 non-null object
date_first_order                          95557 non-null datetime64[ns]
year_first_order                          95557 non-null int64
month_first_order                         95557 non-null int64
total_orders_first_6_months               95557 non-null int64
total_paid_first_6_months                 95557 non-null float64
order_count_unavailable                   95557 non-null int64
avg_payment_processing_time               95557 non-null timedelta64[ns]
avg_seller_processing_time                95557 non-null timedelta64[ns]
avg_transit_time                          95557 non-null timedelta64[ns]
avg_lead_time                             95557 non-null timedelta64[ns]
avg_item_count_per_order                  95557 non-null float64
avg_product_count_per_order               95557 non-null float64
orders_shipped_late                       95557 non-null float64
avg_quantity_by_product                   95557 non-null object
average_price_per_unit                    95557 non-null float64
avg_freight_cost_per_order                95557 non-null float64
ordered_from_top_10_prod_category_bol     95557 non-null int64
ordered_from_seller_perfect_avg_review_bol 95557 non-null int64
avg_survey_lag                            95557 non-null timedelta64[ns]
avg_review_lag                            95557 non-null timedelta64[ns]
avg_review_score                          95557 non-null float64
order_count_boleto_voucher                95557 non-null float64
order_count_card                          95557 non-null float64
avg_installments                          95557 non-null float64
avg_days_payment_processing_time          95557 non-null float64
avg_days_seller_processing_time           95557 non-null float64
avg_days_transit_time                     95557 non-null float64
avg_days_lead_time                        95557 non-null float64
avg_days_survey_lag                       95557 non-null float64
avg_daysreview_lag                        95557 non-null float64
Active?                                   95557 non-null int32
perc_orders_unavailable                   95557 non-null float64
perc_orders_shipped_late                  95557 non-null float64
perc_orders_boleto_voucher                95557 non-null float64
perc_orders_credit_debit                  95557 non-null float64
day_first_order                           95557 non-null object
dtypes: datetime64[ns](1), float64(20), int32(1), int64(6), object(3), timedelta64[ns](6)
memory usage: 26.6+ MB
```

- Replace or drop nulls & update data types
- Columns with counts turned to proportion of total orders to fairly compare customers

VARIABLES AVAILABLE (36)

CUSTOMER BEHAVIOR

customer_unique_id
date_first_order
year_first_order
month_first_order
total_orders_first_6_months
avg_review_lag
total_paid_first_6_months
avg_quantity_by_product
avg_item_count_per_order
avg_product_count_per_order
average_price_per_unit
ordered_from_top_10_prod_category_bol
ordered_from_seller_perfect_avg_review_bol
avg_review_score
order_count_boleto_voucher
order_count_card
avg_installments
perc_orders_boleto/voucher
perc_orders_credit/debit
Active?

LOGISTICS

order_count_unavailable
avg_payment_processing_time
avg_seller_processing_time
avg_transit_time
avg_lead_time
orders_shipped_late
avg_freight_cost_per_order
avg_survey_lag
avg_days_payment_processing_time
avg_days_seller_processing_time
avg_days_transit_time
avg_days_lead_time
avg_days_survey_lag
perc_orders_unavailable
perc_orders_shipped_late

18 Highlighted = Predictors for Model

EXPLORATORY DATA ANALYSIS

What variables do we have now?

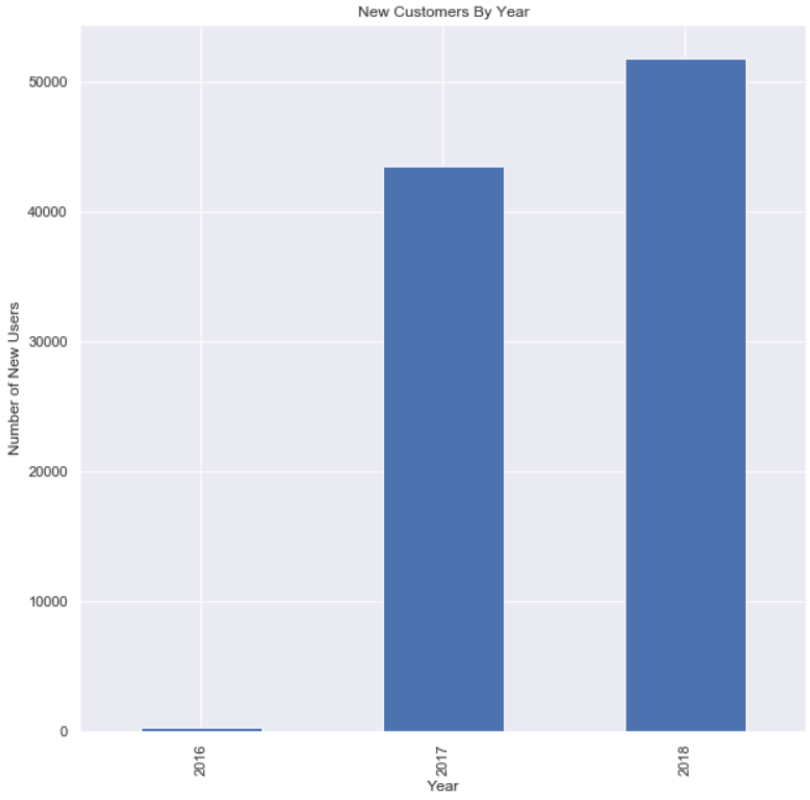
What part of the customer journey do variables fall into?

What insights can we derive for Olist?

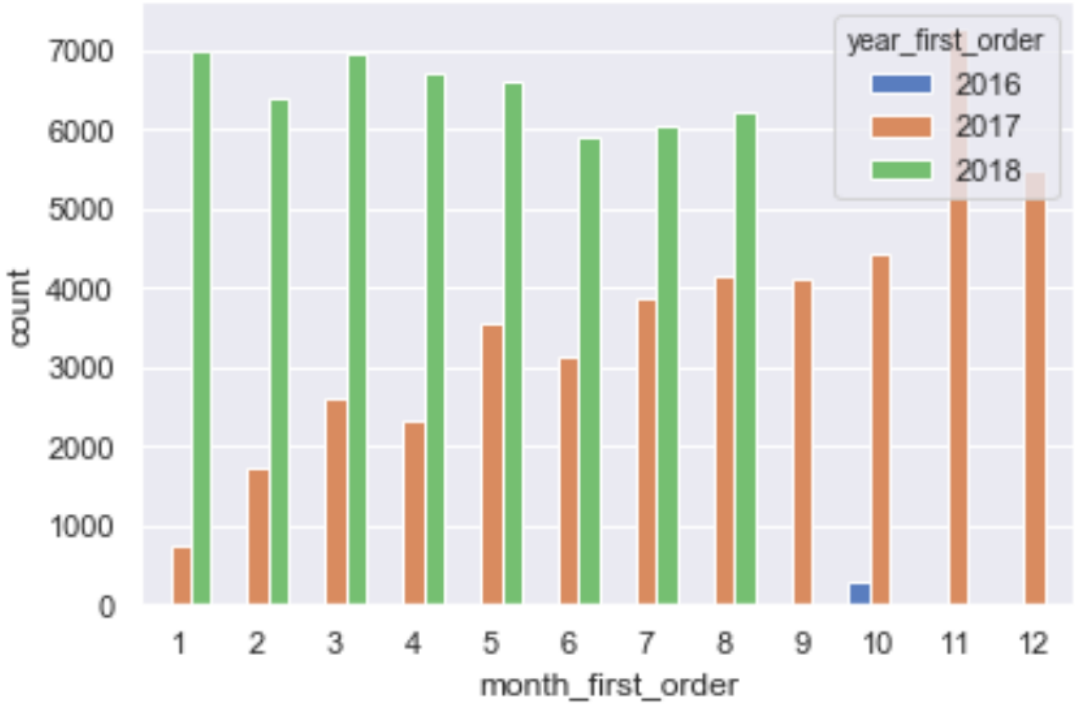
BIG PICTURE

NEW CUSTOMERS

Number of New Customers YOY



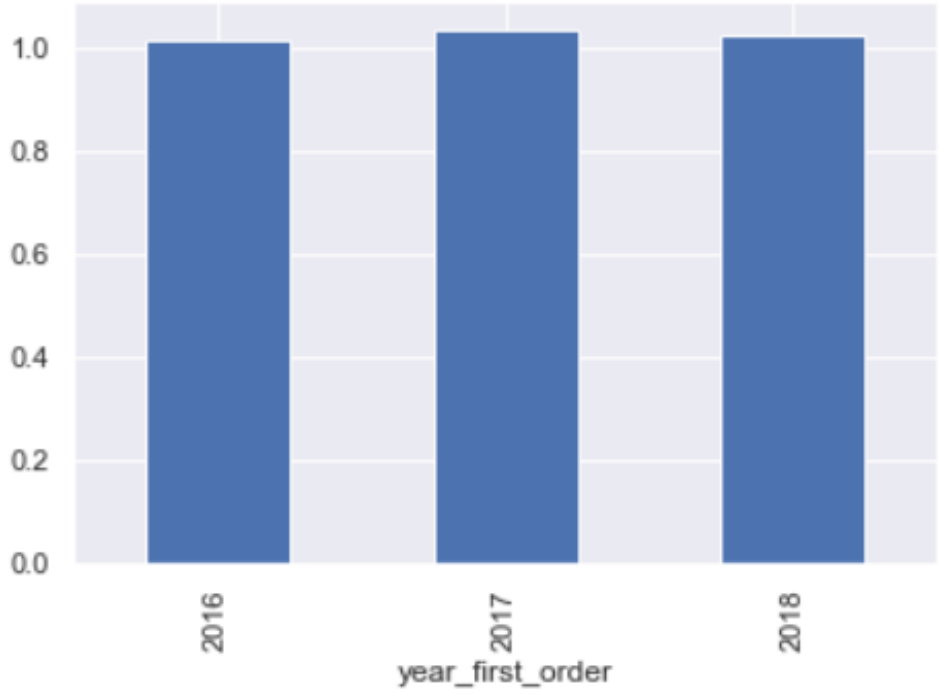
Number of New Customers YOY, broken down by month



- Number of new customers steadily **increasing** per year
- Positive trend **stopped** in 2018, now customer growth stagnant

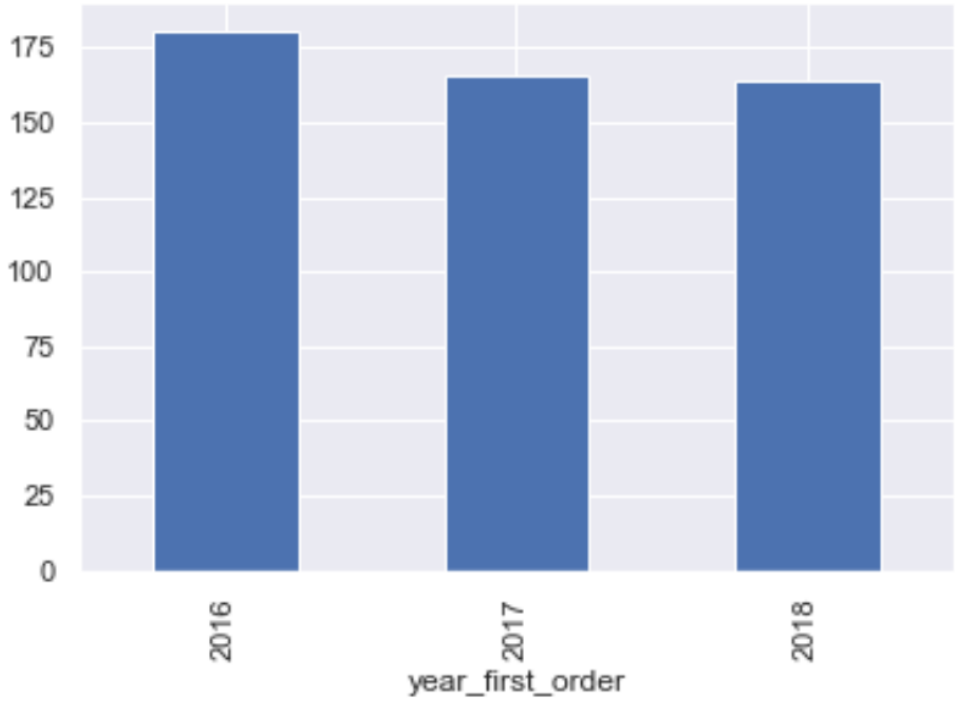
ORDERS PER CUSTOMER

Average number of orders per customer YOY



TOTAL SPENT FIRST 6 MONTHS (LTV)

Average total spent first 6 months YOY

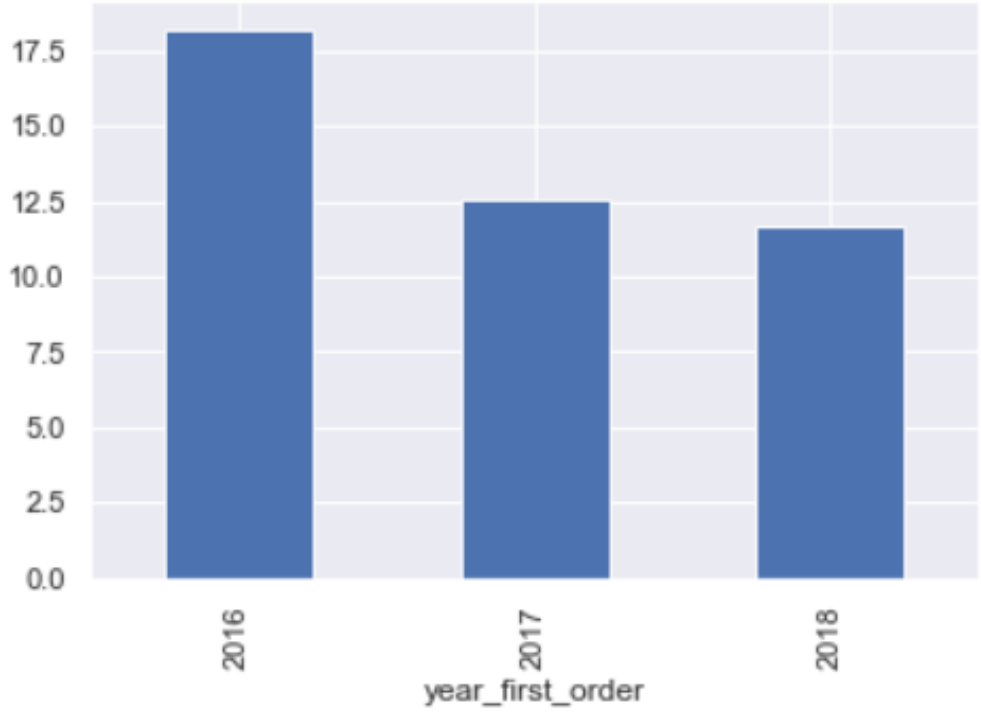


- Number of orders per customer **stagnant** at **1** order per customer. Over **97%** of customers only make 1 order.
- Total spent first 6 month **slightly decreasing** year-on-year

CUSTOMER BEHAVIOR

REVIEWS

Days for customer to submit review, YOY



Average Review Score (Scale 1-5)

4.08

- Year over year, customers are responding to our survey **faster**
- Average review score is **high** and continues to trend positively

PRODUCT CATEGORIES

Top 10 Popular Categories (by customer count)

Rank	Category	% Customers
1	bed_bath_table	9.50%
2	health_beauty	9.00%
3	sports_leisure	7.78%
4	computers_accessories	6.80%
5	furniture_decor	6.56%
6	housewares	6.02%
7	watches_gifts	5.75%
8	telephony	4.32%
9	auto	3.99%
10	toys	3.98%
	Total	**63.69%**

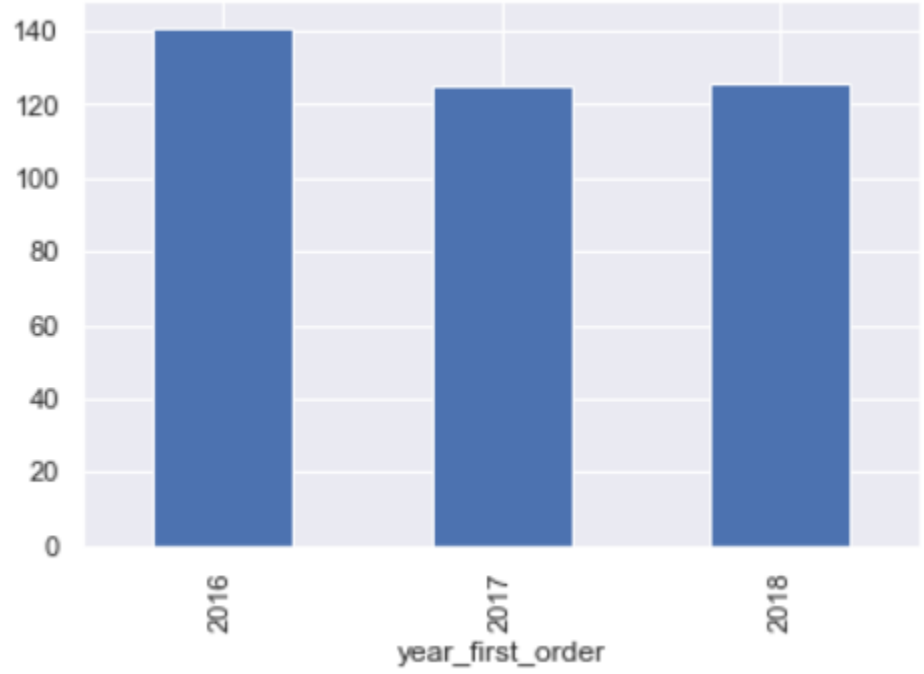
- Over **63%** of customers buy from top 10 categories (out of 71).
- Customers tend to get **1 item per order.**
- Price per unit has been **decreasing** YOY ($15 down from 2016)

ORDER INCLUSIONS

of items per order, by customer

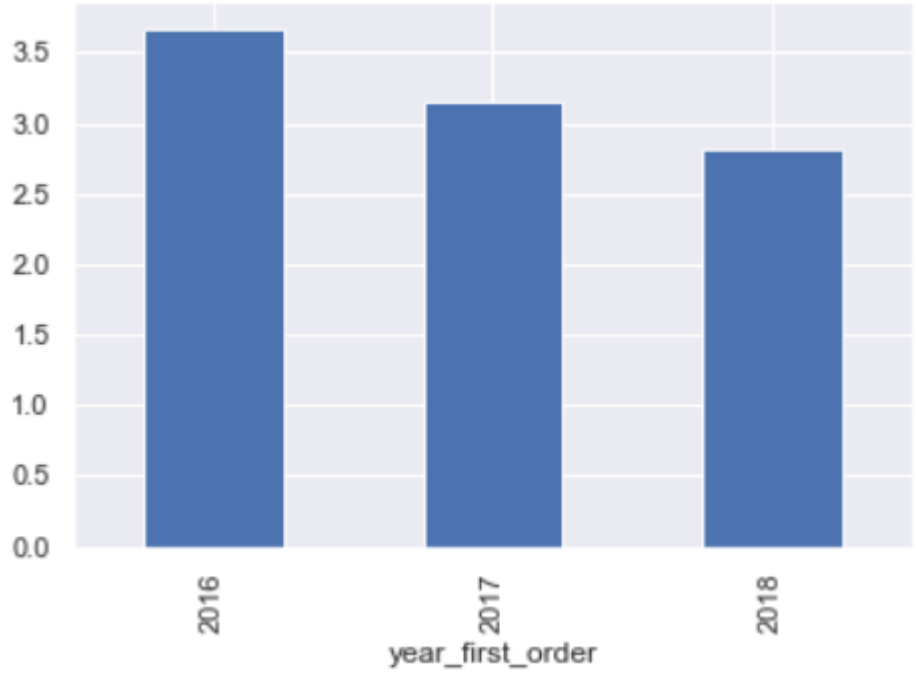
1.14

Price Per Unit, YOY



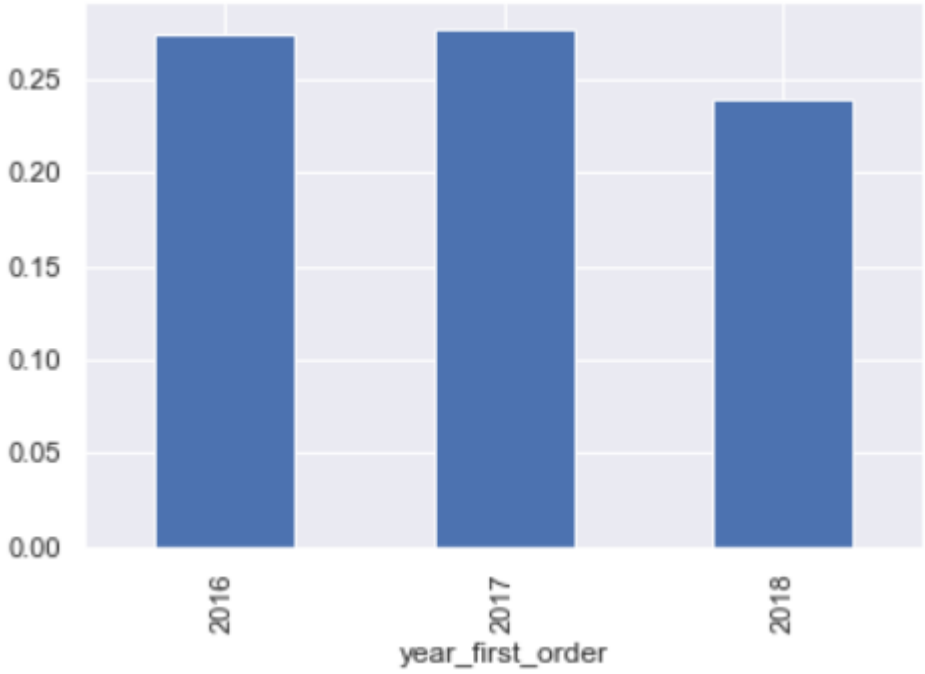
NUMBER OF INSTALLMENTS

Average payment installments per customers YOY



ORDERS PAID BY BOLETO/VOUCHER

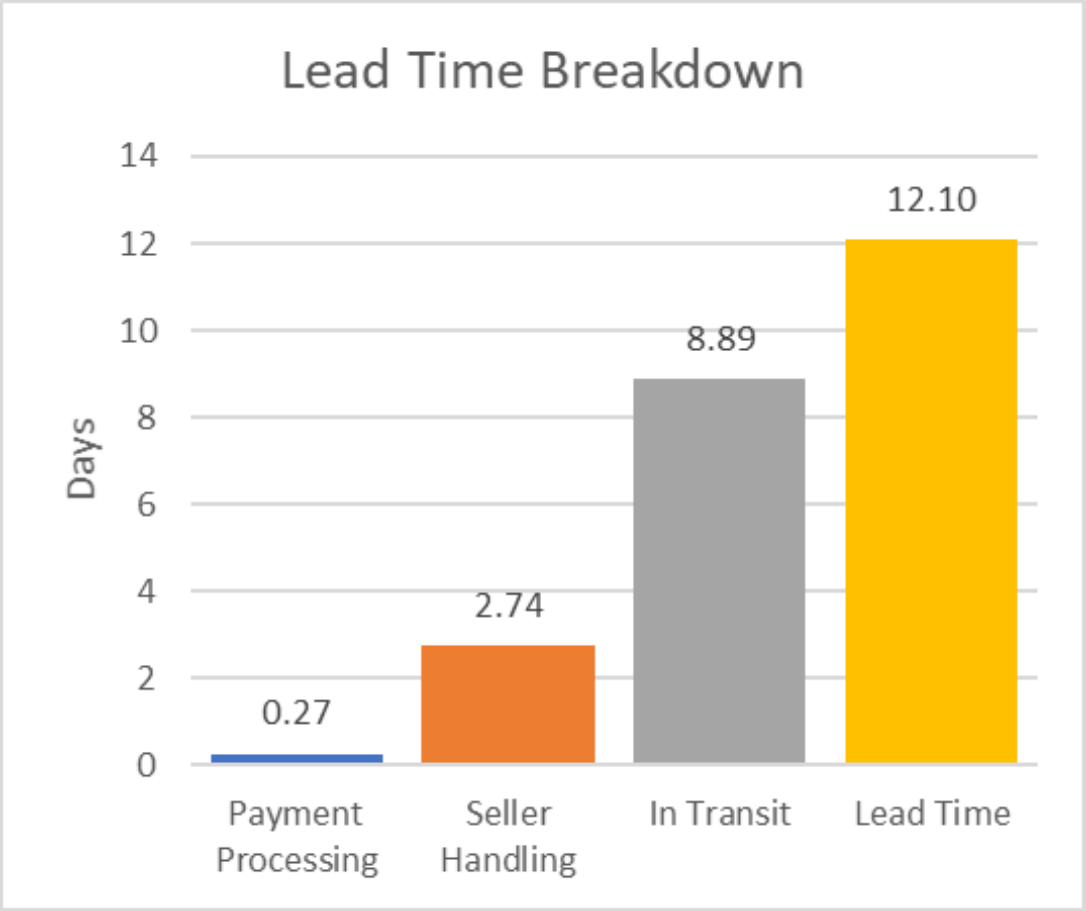
Grouped by customer, proportion of orders paid by boleto or voucher



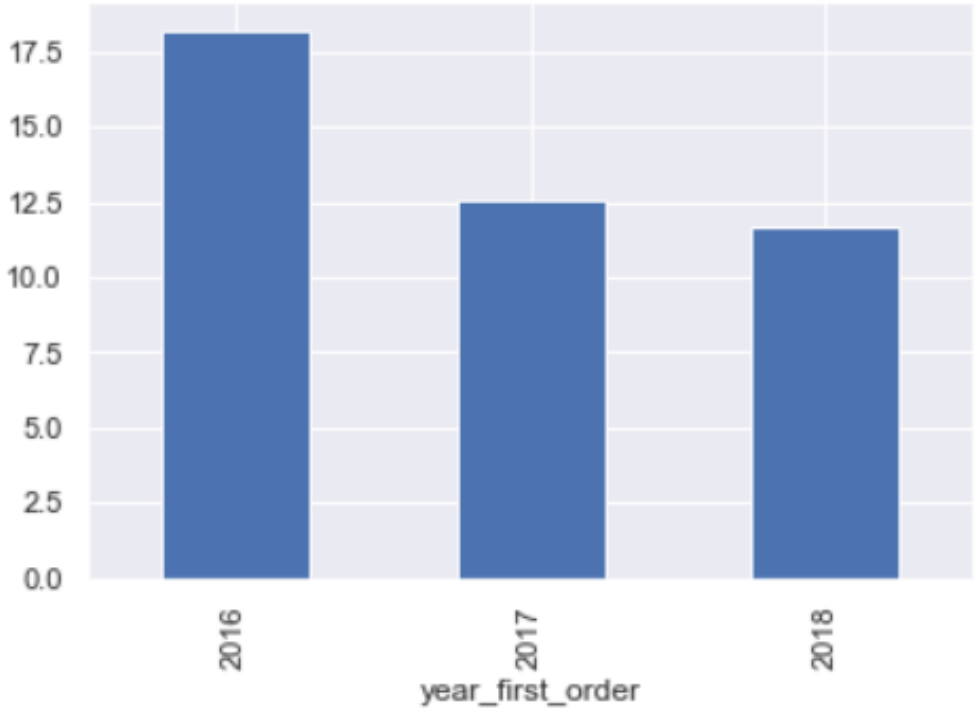
- On average, customers are **decreasing** the number of payment installments
- The proportion of orders paid by boleto or voucher is relatively stable (avg: 25.62%)

LOGISTICS

LEAD TIME



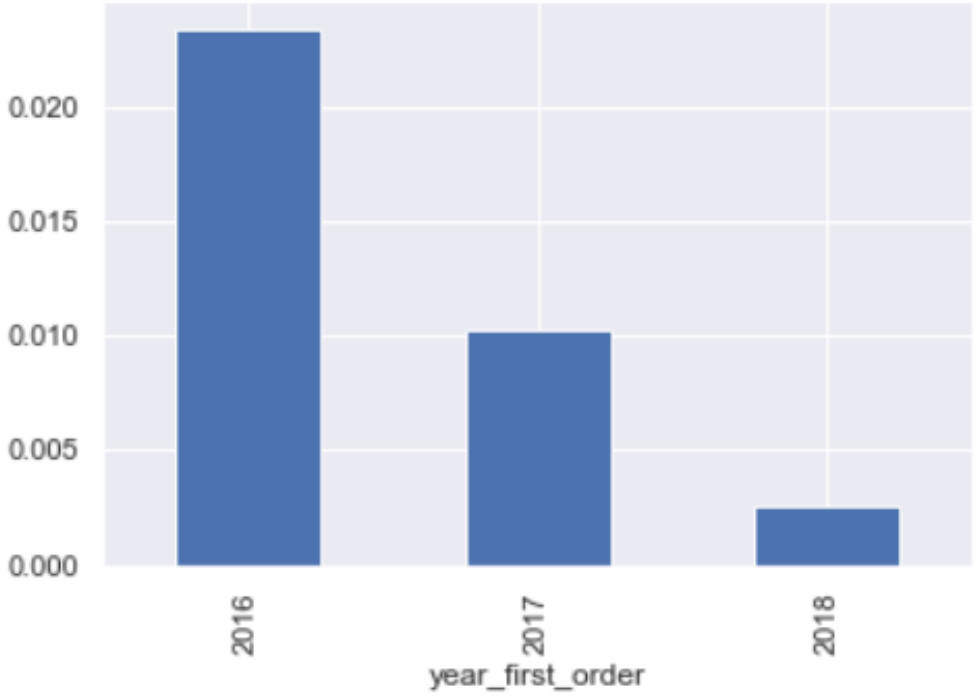
Lead Time (year-over-year)



- The expected lead time for a customer has been **decreasing** year-over-year.
- **Positive trend consistent** across all stages (payment processing, seller handling, and in transit)

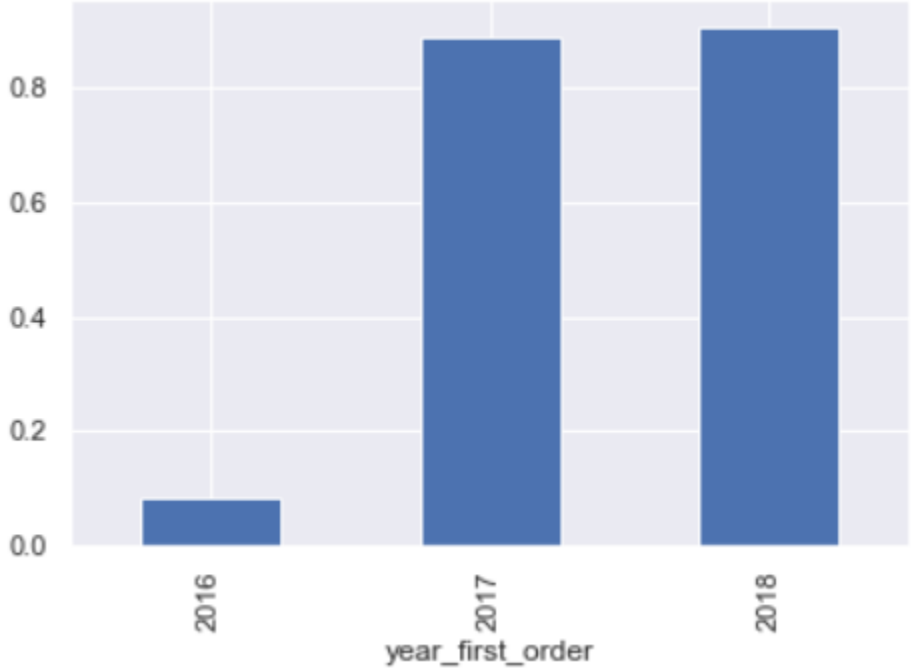
ORDERS UNAVAILABLE

Average proportion of orders unavailable YOY



ORDER SHIPPED LATE

Average proportion of orders unavailable YOY



- The proportion of orders for one customer that will be unavailable has been **decreasing** year over year. Overall: **0.61%**
- The proportion of orders for one customer that will be shipped late has been **increasing** year over year. Overall: **89.7%.** So far, trend has not impacted overall lead_time

REGRESSION MODEL - LTV

MODEL USED

- Statsmodel **Ordinary Least Squares (OLS)** Regression
- Model to **explain** correlation between variables and 6-month customer's lifetime value

RESPONSE VARIABLE

total_paid_first_6_months

PREDICTOR VARIABLES (17)

month_first_order (C)	avg_freight_cost_per_order
ordered_from_top_10_category_bol (C)	avg_installments
ordered_from_seller_perfect_avg_reviews bol (C)	perc_orders_shipped_late
perc_orders_unavailable	perc_orders_boleto_voucher
avg_review_score	avg_days_payment_processing_time
avg_item_count_per_order	avg_days_seller_processing_time
avg_product_count_per_order	avg_days_transit_time
average_price_per_unit	avg_days_survey_lag
	avg_daysreview_lag

MODEL #1

- **22** Variables (including "dummy variables" from categorical data)
- R-squared = **0.877**
- **Missing** timedelta variables
- **4** insignificant variables (P-Score < 0.05)

Dep. Variable:	total_paid_first_6_months	R-squared:	0.877
Model:	OLS	Adj. R-squared:	0.877
Method:	Least Squares	F-statistic:	3.089e+04
Date:	Sun, 08 Dec 2019	Prob (F-statistic):	0.00
Time:	20:17:39	Log-Likelihood:	-5.5397e+05
No. Observations:	95557	AIC:	1.108e+06
Df Residuals:	95534	BIC:	1.108e+06
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-89.2818	1.945	-45.903	0.000	-93.094	-85.470
C(month_first_order)[T.2]	-0.0388	1.267	-0.031	0.976	-2.522	2.444
C(month_first_order)[T.3]	0.1672	1.220	0.137	0.891	-2.225	2.559
C(month_first_order)[T.4]	-0.4776	1.235	-0.387	0.699	-2.899	1.944
C(month_first_order)[T.5]	1.0265	1.203	0.853	0.394	-1.332	3.385
C(month_first_order)[T.6]	-2.0731	1.236	-1.677	0.094	-4.496	0.350
C(month_first_order)[T.7]	-0.4061	1.211	-0.335	0.737	-2.780	1.967
C(month_first_order)[T.8]	-1.5415	1.198	-1.286	0.198	-3.890	0.807
C(month_first_order)[T.9]	5.8527	1.539	3.803	0.000	2.836	8.869
C(month_first_order)[T.10]	-1.8868	1.470	-1.283	0.199	-4.768	0.995
C(month_first_order)[T.11]	0.2644	1.302	0.203	0.839	-2.288	2.817
C(month_first_order)[T.12]	-0.1119	1.407	-0.080	0.937	-2.870	2.646
C(ordered_from_top_10_prod_category_bol)[T.1]	2.2730	0.540	4.213	0.000	1.216	3.330
C(ordered_from_seller_perfect_avg_review_bol)[T.1]	-1.2899	2.627	-0.491	0.623	-6.440	3.860
perc_orders_unavailable	177.5627	3.456	51.376	0.000	170.789	184.337
avg_review_score	-0.2671	0.200	-1.337	0.181	-0.659	0.124
avg_item_count_per_order	104.7951	0.560	187.250	0.000	103.698	105.892
avg_product_count_per_order	-24.0295	1.368	-17.566	0.000	-26.711	-21.348
average_price_per_unit	1.0420	0.002	673.569	0.000	1.039	1.045
avg_freight_cost_per_order	1.2655	0.017	73.331	0.000	1.232	1.299
avg_installments	0.7602	0.103	7.348	0.000	0.557	0.963
perc_orders_shipped_late	-1.3082	0.891	-1.468	0.142	-3.054	0.438
perc_orders_boleto_voucher	1.5403	0.475	3.245	0.001	0.610	2.471

Omnibus:	283960.659	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	81936180578.570
Skew:	42.153	Prob(JB):	0.00
Kurtosis:	4538.626	Cond. No.	3.10e+03

MODEL OPTIMIZATION

Model #	Change Description	N. Of Variables (including dummy)	R-Squared	N. of Variables with p-square > 0.05
1		22	0.877	4
2	Adds timedelta variables as n. of days	28	0.877	6
3	Removes month_first_order (C)	17	0.877	5
4	Removes perc_orders_shipped_late	16	0.877	4
5	Removes ordered_from_seller_perfect_review (C)	15	0.877	3
6	Removes avg_days_payment_processing_time	14	0.877	2
7	Removes avg_daysreview_lag	13	0.877	1
8	Removes avg_review_score	12	0.877	0
9	Normalizes data (z-score)	12	0.877	0

FINAL MODEL

- **Positive** Correlation to LTV
  - **ordered from top 10 categories**
  - % orders unavailable
  - Item count per order
  - price per unit
  - freight cost per order
  - **# payment installments**
  - **% orders paid boleto or voucher**
  - Survey Lag (from Olist to customer)

- **Negative** Correlation to LTV
  - product count per order
  - **seller processing time**
  - **transit time**

OLS Regression Results

Dep. Variable:	total_paid_first_6_months	**R-squared:**	0.877
Model:	OLS	**Adj. R-squared:**	0.877
Method:	Least Squares	**F-statistic:**	5.664e+04
Date:	Mon, 09 Dec 2019	**Prob (F-statistic):**	0.00
Time:	02:19:32	**Log-Likelihood:**	-5.5396e+05
No. Observations:	95557	**AIC:**	1.108e+06
Df Residuals:	95544	**BIC:**	1.108e+06
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-92.7899	1.372	-67.611	0.000	-95.480	-90.100
C(ordered_from_top_10_prod_category_bol)[T.1]	2.3119	0.539	4.288	0.000	1.255	3.369
perc_orders_unavailable	179.6209	3.338	53.816	0.000	173.079	186.163
avg_item_count_per_order	104.7614	0.558	187.655	0.000	103.667	105.856
avg_product_count_per_order	-23.8962	1.369	-17.454	0.000	-26.580	-21.213
average_price_per_unit	1.0418	0.002	672.494	0.000	1.039	1.045
avg_freight_cost_per_order	1.2673	0.018	72.197	0.000	1.233	1.302
avg_installments	0.7760	0.103	7.504	0.000	0.573	0.979
perc_orders_boleto_voucher	1.4128	0.476	2.965	0.003	0.479	2.347
avg_days_seller_processing_time	-1.0578	0.281	-3.766	0.000	-1.608	-0.507
avg_days_transit_time	-1.3576	0.278	-4.879	0.000	-1.903	-0.812
avg_days_lead_time	1.3025	0.278	4.691	0.000	0.758	1.847
avg_days_survey_lag	1.0362	0.214	4.850	0.000	0.617	1.455

Omnibus:	284112.843	**Durbin-Watson:**	2.004
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	82359323782.783
Skew:	42.215	**Prob(JB):**	0.00
Kurtosis:	4550.324	**Cond. No.**	2.97e+03

POTENTIAL NEXT STEPS FOR PROJECT

- Churn Rate or Survival Analysis
- Customer Segmentation
- Logistics Audit
- Content Review (NLP)
- Sentiment Analysis
- Seller Patterns
- Inventory Review

TOOLS USED:

- Postgres SQL
- Python:
  Pandas, Numpy, Matplotlib, Statsmodel, Seaborn, Scipy

MORE INFORMATION ON THIS PROJECT:

- github.com/rachelleaperez
- linkedin.com/in/rachelleperez/

THANK YOU