olist

Final Project
Olist Data Analysis (2016-2018)

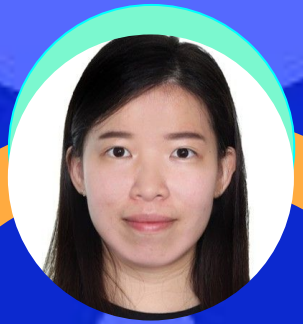

Saifuddin


Ashleigh Wang


Pierre Lim


Quek Kai Ying


Dorothy Ng


Choong Sook Yin

Data Analyst

Data Engineer

olist

Agenda

- Background
- ETL Process:
- Extract (Schema), Transform (Cleaning on Python and Power BI), Load(On PostgreSQL)
- Problem Statement
- **Data exploration - Any generic trends / descriptive stats (Power BI)**
- **Data Visualisation on Product & Reviews & Business Recommendation 1 (Power BI)**
- **Data Visualisation on Logistics & Business Recommendation 2 (Power BI)**
- **Data Visualisation on Customers & Business Recommendation 3 (Power BI)**
- Recommendations summary
- Limitations of data

Background

E-COMMERCE MARKET ANALYSIS

The eCommerce market in Brazil

Brazil is the 15th largest market for eCommerce with a revenue of US$22 billion in 2020, placing it ahead of Netherlands and behind Italy.

Brazil e-commerce to grow 31% in 2021 to US$35bn – Goldman Sachs

From 2021 to 2024, the bank expects e-commerce in Brazil to grow an average of 25% per year, taking the online share to 20.2% of total retail in 2024, from 11.1% in 2020.

By **The Rio Times** - October 22, 2021

RIO DE JANEIRO, BRAZIL - Goldman Sachs estimates that Brazilian e-commerce as a whole should grow 31% in 2021, reaching R$201 (US$35) billion in gross sales (GMV) and implying a 13% penetration in the total retail market, up 1.95% year over year.

The new wave of Brazilian SaaS innovators

Diego Gomes 5:00 AM GMT+8 • August 9, 2017 Comment

ETL Process

Extract

Read data from CSV files

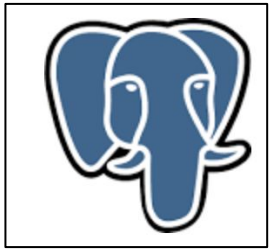
Messy data storage with scattered data in multiple tables

Duplicated data

Incorrect or NULL values

Transform

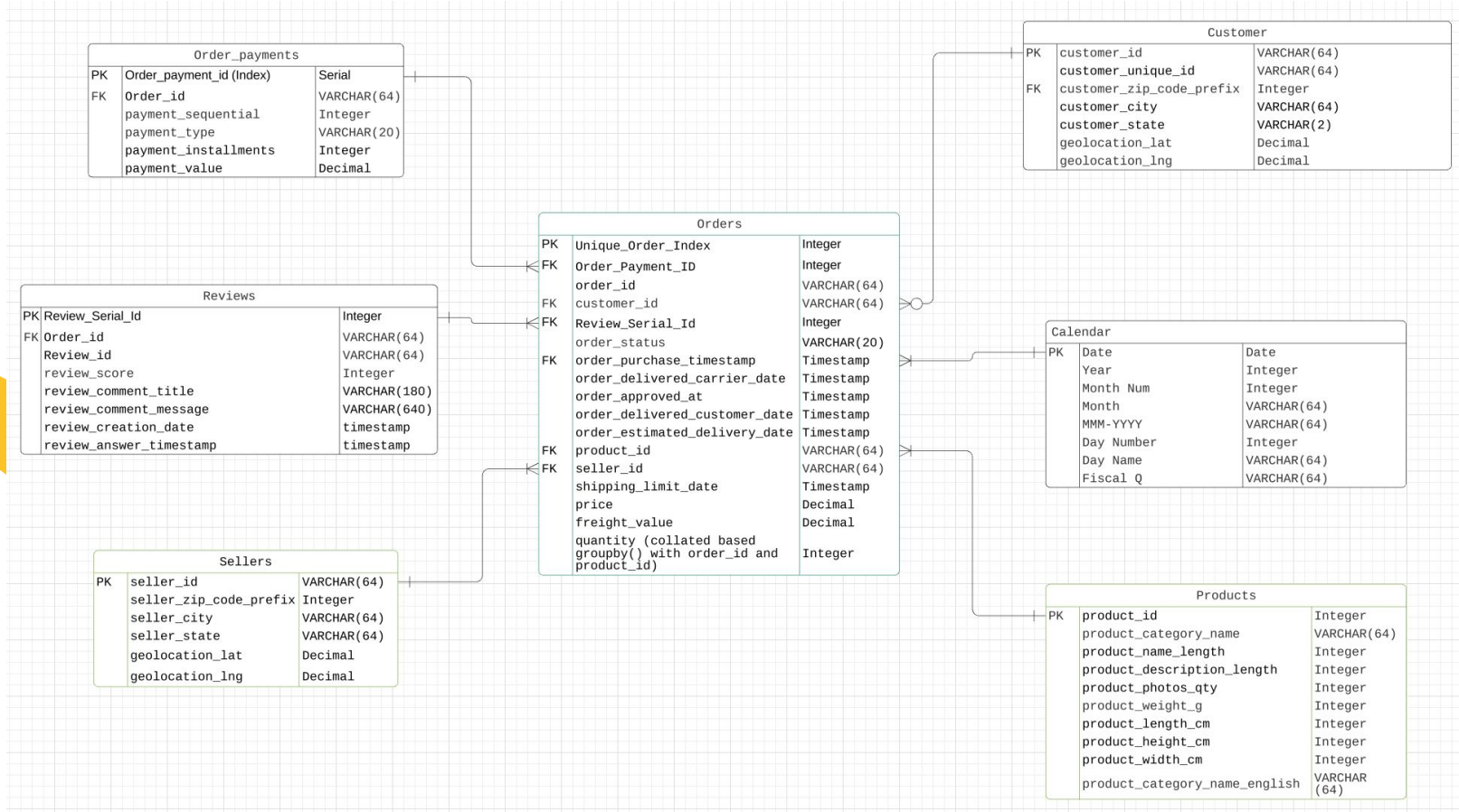
Data cleaning and joining of tables

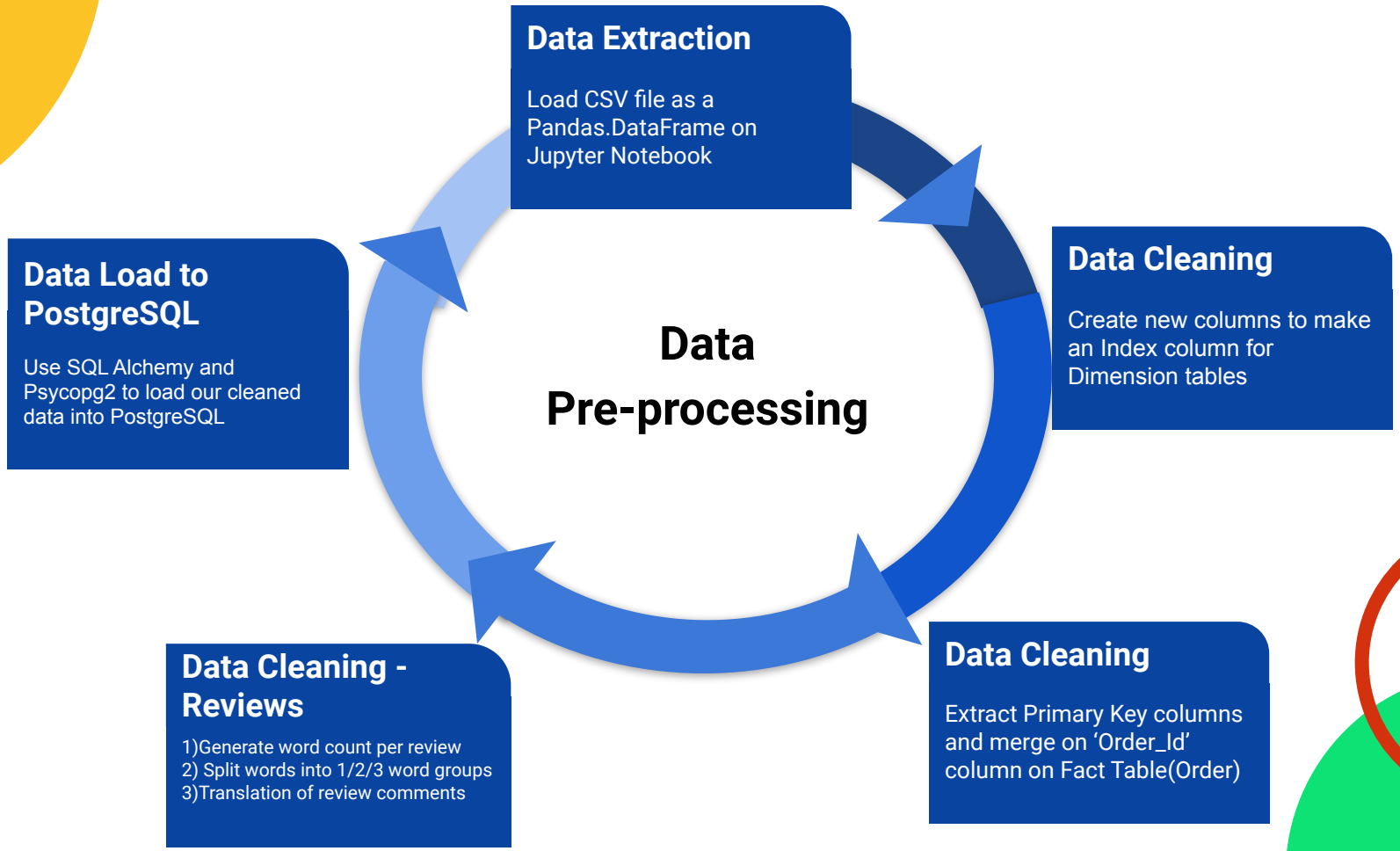


Load

Store clean data in database

Schema


Order_payments
PK Order_payment_id (Index) Serial
FK Order_id VARCHAR(64)
payment_sequential Integer
payment_type VARCHAR(20)
payment_installments Integer
payment_value Decimal
Customer
PK customer_id VARCHAR(64)
customer_unique_id VARCHAR(64)
FK customer_zip_code_prefix Integer
customer_city VARCHAR(64)
customer_state VARCHAR(2)
geolocation_lat Decimal
geolocation_lng Decimal
Orders
PK Unique_Order_Index Integer
FK Order_Payment_ID Integer
order_id VARCHAR(64)
FK customer_id VARCHAR(64)
FK Review_Serial_Id Integer
order_status VARCHAR(20)
FK order_purchase_timestamp Timestamp
order_delivered_carrier_date Timestamp
order_approved_at Timestamp
order_delivered_customer_date Timestamp
order_estimated_delivery_date Timestamp
FK product_id VARCHAR(64)
FK seller_id VARCHAR(64)
shipping_limit_date Timestamp
price Decimal
freight_value Decimal
quantity (collated based groupby() with order_id and product_id) Integer
Reviews
PK Review_Serial_Id Integer
FK Order_id VARCHAR(64)
Review_id VARCHAR(64)
review_score Integer
review_comment_title VARCHAR(180)
review_comment_message VARCHAR(640)
review_creation_date timestamp
review_answer_timestamp timestamp
Calendar
PK Date Date
Year Integer
Month Num Integer
Month VARCHAR(64)
MMM-YYYY VARCHAR(64)
Day Number Integer
Day Name VARCHAR(64)
Fiscal Q VARCHAR(64)
Sellers
PK seller_id VARCHAR(64)
seller_zip_code_prefix Integer
seller_city VARCHAR(64)
seller_state VARCHAR(64)
geolocation_lat Decimal
geolocation_lng Decimal
Products
PK product_id Integer
product_category_name VARCHAR(64)
product_name_length Integer
product_description_length Integer
product_photos_qty Integer
product_weight_g Integer
product_length_cm Integer
product_height_cm Integer
product_width_cm Integer
product_category_name_english VARCHAR(64)


Data Extraction
Load CSV file as a Pandas.DataFrame on Jupyter Notebook
Data Cleaning
Create new columns to make an Index column for Dimension tables
Data Cleaning
Extract Primary Key columns and merge on 'Order_Id' column on Fact Table(Order)
Data Cleaning - Reviews
1)Generate word count per review
2) Split words into 1/2/3 word groups
3)Translation of review comments
Data Load to PostgreSQL
Use SQL Alchemy and Psycopg2 to load our cleaned data into PostgreSQL
Data Pre-processing

Data Cleaning

In [5]:
```
#normalize special characters in portuguese to english characters using unidecode
#change null values in review comment message to NaN

reviews["review_comment_message"] = reviews["review_comment_message"].apply(unidecode)
reviews.loc[reviews['review_comment_message']== 'nan', 'review_comment_message'] = np.nan

reviews
```

Out[5]:

	review_id	order_id	review_score	review_comment_title	review_comment_message	review_creation_
0	7bc2406110b926393aa56f80a40eba40	73fc7af87114b39712e6da79b0a377eb	4	NaN	NaN	2018-01-18 00:0
1	80e641a11e56f04c1ad469d5645fdfde	a548910a1c6147796b98fdf73dbeba33	5	NaN	NaN	2018-03-10 00:0
2	228ce5500dc1d8e020d8d1322874b6f0	f9e4b658b201a9f2ecdecbb34bed034b	5	NaN	NaN	2018-02-17 00:0
3	e64fb393e7b32834bb789ff8bb30750e	658677c97b385a9be170737859d3511b	5	NaN	Recebi bem antes do prazo estipulado.	2017-04-21 00:0
4	f7c4243c7fe1938f181bec41a392bdeb	8e6bfb81e283fa7e4f11123a3fb894f1	5	NaN	Parabens lojas lannister adorei comprar pela I...	2018-03-01 00:0
...	...	...	...	...	...	
99995	f3897127253a9592a73be9bdfdf4ed7a	22ec9f0669f784db00fa86d035cf8602	5	NaN	NaN	2017-12-09 00:0
99996	b3de70c89b1510c4cd3d0649fd302472	55d4004744368f5571d1f590031933e4	5	NaN	Excelente mochila, entrega super rapida. Super...	2018-03-22 00:0
99997	1adeb9d84d72fe4e337617733eb85149	7725825d039fc1f0ceb7635e3f7d9206	4	NaN	NaN	2018-07-01 00:0
99998	be360f18f5df1e0541061c87021e6d93	f8bd3f2000c28c5342fedeb5e50f2e75	1	NaN	Solicitei a compra de uma capa de retrovisor c...	2017-12-15 00:0
99999	efe49f1d6f951dd88b51e6ccd4cc548f	90531360ecb1eec2a1fbb265a0db0508	1	NaN	meu produto chegou e ja tenho que devolver, po...	2017-07-03 00:0

100000 rows × 8 columns

Data Cleaning

In [6]:
```
#remove rows with no comments

reviews_withoutNA = reviews[reviews['review_comment_message'].notna()]

reviews_withoutNA
```

Out[6]:

	review_id	order_id	review_score	review_comment_title	review_comment_message	review_creation_
3	e64fb393e7b32834bb789ff8bb30750e	658677c97b385a9be170737859d3511b	5	NaN	Recebi bem antes do prazo estipulado.	2017-04-21 00:0
4	f7c4243c7fe1938f181bec41a392bdeb	8e6bfb81e283fa7e4f11123a3fb894f1	5	NaN	Parabens lojas lannister adorei comprar pela I...	2018-03-01 00:0
9	8670d52e15e00043ae7de4c01cc2fe06	b9bf720beb4ab3728760088589c62129	4	recomendo	aparelho eficiente. no site a marca do aparelh...	2018-05-22 00:0
12	4b49719c8a200003f700d3d986ea1a19	9d6f15f95d01e79bd1349cc208361f09	4	NaN	Mas um pouco ,travando...pelo valor ta Boa.\r\n	2018-02-16 00:0
15	3948b09f7c818e2d86c9a546758b2335	e51478e7e277a83743b6f9991dbfa3fb	5	Super recomendo	Vendedor confiavel, produto ok e entrega antes...	2018-05-23 00:0
...	...	...	...	...	...	
99983	df5fae90e85354241d5d64a8955b2b09	509b86c65fe4e2ad5b96408cfef9755e	5	NaN	Entregou dentro do prazo. O produto chegou em ...	2018-02-07 00:0
99990	a709d176f59bc3af77f4149c96bae357	d5cb12269711bd1eaf7eed8fd32a7c95	3	NaN	O produto nao foi enviado com NF, nao existe v...	2018-05-19 00:0
99996	b3de70c89b1510c4cd3d0649fd302472	55d4004744368f5571d1f590031933e4	5	NaN	Excelente mochila, entrega super rapida. Super...	2018-03-22 00:0
99998	be360f18f5df1e0541061c87021e6d93	f8bd3f2000c28c5342fedeb5e50f2e75	1	NaN	Solicitei a compra de uma capa de retrovisor c...	2017-12-15 00:0
99999	efe49f1d6f951dd88b51e6ccd4cc548f	90531360ecb1eec2a1fbb265a0db0508	1	NaN	meu produto chegou e ja tenho que devolver, po...	2017-07-03 00:0

41753 rows × 8 columns

Data Cleaning

In [9]:
```
#add column to count number of words in each review comment
#will be exported as csv for powerbi viz

reviews_withoutNA['word_count'] = reviews_withoutNA.review_comment_message.apply(lambda x: len(str(x).split()))

#reviews.to_csv('reviews_withoutNA')
reviews_withoutNA
```

```
<ipython-input-9-d5e1defc3e27>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  reviews_withoutNA['word_count'] = reviews_withoutNA.review_comment_message.apply(lambda x: len(str(x).split()))
```

Out[9]:

order_id	review_score	review_comment_title	review_comment_message	review_creation_date	review_answer_timestamp	review_serial_id	word_count
be170737859d3511b	5	NaN	Recebi bem antes do prazo estipulado.	2017-04-21 00:00:00	2017-04-21 22:02:06	4	6
a7e4f11123a3fb894f1	5	NaN	Parabens lojas lannister adorei comprar pela I...	2018-03-01 00:00:00	2018-03-02 10:26:53	5	15
28760088589c62129	4	recomendo	aparelho eficiente. no site a marca do aparelh...	2018-05-22 00:00:00	2018-05-23 16:45:47	10	30
9bd1349cc208361f09	4	NaN	Mas um pouco ,travando...pelo valor ta Boa.\r\n	2018-02-16 00:00:00	2018-02-20 10:52:22	13	7
3743b6f9991dbfa3fb	5	Super recomendo	Vendedor confiavel, produto ok e entrega antes...	2018-05-23 00:00:00	2018-05-24 03:00:01	16	9
...	...	...	...	...	...	...	...
ad5b96408cfef9755e	5	NaN	Entregou dentro do prazo. O produto chegou em ...	2018-02-07 00:00:00	2018-02-19 19:47:23	99984	13
1eaf7eed8fd32a7c95	3	NaN	O produto nao foi enviado com NF, nao existe v...	2018-05-19 00:00:00	2018-05-20 21:51:06	99991	25
5571d1f590031933e4	5	NaN	Excelente mochila, entrega super rapida. Super...	2018-03-22 00:00:00	2018-03-23 09:10:43	99997	9
5342fedeb5e50f2e75	1	NaN	Solicitei a compra de uma capa de retrovisor c...	2017-12-15 00:00:00	2017-12-16 01:29:43	99999	33
c2a1fbb265a0db0508	1	NaN	meu produto chegou e ja tenho que devolver, po...	2017-07-03 00:00:00	2017-07-03 21:01:49	100000	16

Data Cleaning

In [11]:
```
#Remove non-alphanumeric characters in review comments
#select relevant columns

reviews_withoutNA['review_comment_message'].replace(regex=True, inplace=True, to_replace=r'[^a-zA-Z\s]+', value=r' ')

review_comments = reviews_withoutNA[["review_score", "review_comment_message", "review_serial_id"]]
review_comments.head(20)
```

Out[11]:

	review_score	review_comment_message	review_serial_id
3	5	Recebi bem antes do prazo estipulado	4
4	5	Parabens lojas lannister adorei comprar pela I...	5
9	4	aparelho eficiente no site a marca do aparelh...	10
12	4	Mas um pouco travando pelo valor ta Boa \r\n	13
15	5	Vendedor confiavel produto ok e entrega antes...	16
16	2	GOSTARIA DE SABER O QUE HOUVE SEMPRE RECEBI E...	17
19	1	Pessimo	20
22	5	Loja nota	23
24	5	obrigado pela atencao amim dispensada	25
27	5	A compra foi realizada facilmente \r\nA entreg...	28
28	5	relogio muito bonito e barato	29
29	1	Nao gostei Comprei gato por lebre	30
32	1	Sempre compro pela Internet e a entrega ocorre...	33
34	4	Recebi exatamente o que esperava As demais en...	35
36	5	Recomendo	37
37	5	muito boa	38
38	5	To completamente apaixonada loja super respon...	39
39	1	Nada de chegar o meu pedido	40
43	5	Muito bom muito cheiroso	44
47	5	otimo vendedor chegou ate antes do prazo ado...	48

Data Cleaning

In [40]:
```
#filter out 1-2 star reviews

bad_reviews = review_comments.loc[review_comments['review_score'].isin([1, 2])]


#filter out 5 star reviews

good_reviews = review_comments.loc[review_comments['review_score'] == 5]
```

Out[40]:

	review_score	review_comment_message	review_serial_id
3	5	Recebi bem antes do prazo estipulado	4
4	5	Parabens lojas lannister adorei comprar pela I...	5
15	5	Vendedor confiavel produto ok e entrega antes...	16
22	5	Loja nota	23
24	5	obrigado pela atencao amim dispensada	25
...	...	...	...
99966	5	Gostei muito \r\nE a segunda vez que compro o ...	99967
99971	5	Ficamos muito satisfeitos com o produto atend...	99972
99977	5	Produto original prazo de entrega rapido Super...	99978
99983	5	Entregou dentro do prazo O produto chegou em ...	99984
99996	5	Excelente mochila entrega super rapida Super...	99997

20646 rows × 3 columns

In [42]:
```
#import nltk package and stopwords package
#create variable with all Portuguese stopwords

import nltk
from nltk.corpus import stopwords


stop = stopwords.words('portuguese')
```

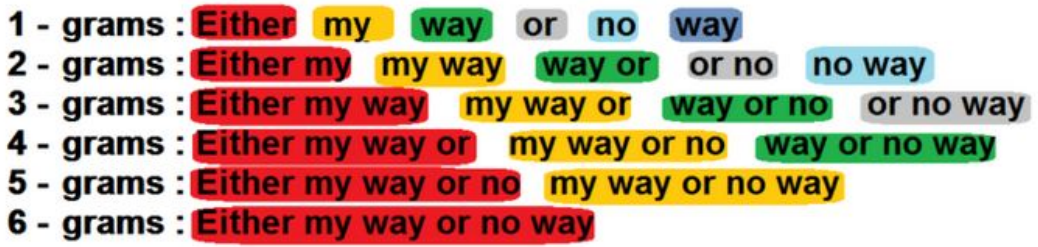
Data Cleaning

In [43]:
```
#generate each comment without stopwords
bad_reviews['comment_without_stopwords'] =
bad_reviews['review_comment_message'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop)]))

#generate unigrams (convert all words to lowercase, split into individual words)
unigrams  = (bad_reviews['comment_without_stopwords'].str.lower()
             .str.split(expand=True)
             .stack())

#generate bigrams (2 word pairs) by concatenating unigram columns
bigrams = unigrams + ' ' + unigrams.shift(-1)
#generate trigrams (3 word pairs) by concatenating unigram and bigram columns
trigrams = bigrams + ' ' + unigrams.shift(-2)

#combine unigram, bigram, trigrams in one dataframe and remove NaNs
combined_bad = pd.concat([unigrams, bigrams, trigrams]).dropna().reset_index(drop=True)
bad_review_ngrams = combined_bad.to_frame().rename(columns={0: 'ngrams'})


a = bad_review_ngrams['ngrams'].value_counts().reset_index()
a['word_count'] = [len(x.split()) for x in a['index'].tolist()]

#exported to csv for powerbi viz
#a.to_csv('bad_reviews.csv')
a
```

Out[43]:

	index	ngrams	word_count
0	nao	8889	1
1	produto	6495	1
2	recebi	3431	1
3	comprei	1982	1
4	nao recebi	1935	2
...	...	...	...
186536	correios rio fazendo	1	3
186537	informando produto comprei	1	3
186538	demora peco	1	2
186539	entregaram como	1	2
186540	ir produto	1	2

186541 rows × 3 columns



Data Cleaning

In [44]:
```
#generate each comment without stopwords
good_reviews['comment_without_stopwords'] =
good_reviews['review_comment_message'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop)]))

#generate unigrams (convert all words to lowercase, split into individual words)
unigrams  = (good_reviews['comment_without_stopwords'].str.lower()
             .str.split(expand=True)
             .stack())

# generate bigrams by concatenating unigram columns
bigrams = unigrams + ' ' + unigrams.shift(-1)
# generate trigrams by concatenating unigram and bigram columns
trigrams = bigrams + ' ' + unigrams.shift(-2)

#combine unigram, bigram, trigrams in one dataframe and remove NaNs
combined_good = pd.concat([unigrams, bigrams, trigrams]).dropna().reset_index(drop=True)
good_review_ngrams = combined_good.to_frame().rename(columns={0: 'ngrams'})


b = good_review_ngrams['ngrams'].value_counts().reset_index()
b['word_count'] = [len(x.split()) for x in b['index'].tolist()]

#exported to csv for powerbi viz
#b.to_csv('good_reviews.csv')
b
```

Out[44]:

	index	ngrams	word_count
0	produto	8167	1
1	prazo	5717	1
2	antes	4537	1
3	entrega	3724	1
4	antes prazo	3362	2
...	...	...	...
135021	melhor sao	1	2
135022	brasil rapido	1	2
135023	material gostei bastante	1	3
135024	ok lindo entrega	1	3
135025	recomendado relogio	1	2

135026 rows × 3 columns

Data Loading

In [28]:
```
orders = orders.set_index('order_index')
sellers = sellers.set_index('seller_id')
customer = customer.set_index('customer_id')
order_items.set_index('transaction_id')
products = products.set_index('product_id')
order_payments = order_payments.set_index('order_payment_id')
reviews = reviews.set_index('review_serial_id')
```

In [45]:
```
from psycopg2.extensions import ISOLATION_LEVEL_AUTOCOMMIT
pgcon = psycopg2.connect(host = 'localhost', user = 'postgres', password = 'bida123')
pgcursor = pgcon.cursor()
pgcon.set_isolation_level(ISOLATION_LEVEL_AUTOCOMMIT)
pgcursor.execute('DROP DATABASE IF EXISTS Final_Project')
pgcursor.execute('CREATE DATABASE Final_Project')
pgcon.close()
pgcon = psycopg2.connect(host = 'localhost', database = 'final_project', user = 'postgres', password = 'bida123')
engine = create_engine('postgresql+psycopg2://postgres:bida123@localhost/final_project')
```

In [46]:
```
customer.to_sql('customer', engine, if_exists='replace', index = True)
geolocation.to_sql('geolocation', engine, if_exists='replace', index = True)
order_items.to_sql('order_items', engine, if_exists='replace', index = True)
order_payments.to_sql('order_payments', engine, if_exists='replace', index = True)
reviews.to_sql('reviews', engine, if_exists='replace', index = True)
orders.to_sql('orders', engine, if_exists='replace', index = True)
products.to_sql('products', engine, if_exists='replace', index = True)
sellers.to_sql('sellers', engine, if_exists='replace', index = True)
product_category_name_translation.to_sql('product_category_name_translation', engine, if_exists='replace', index = True)
```

Data Loading

In [ ]:

```
engine.execute('ALTER TABLE customer ALTER COLUMN customer_zip_code_prefix TYPE int
               USING customer_zip_code_prefix :: integer')
engine.execute('ALTER TABLE customer ALTER COLUMN customer_id TYPE varchar(64)')
engine.execute('ALTER TABLE customer ALTER COLUMN customer_unique_id TYPE varchar(64)')
engine.execute('ALTER TABLE customer ALTER COLUMN customer_city TYPE varchar(64)')
engine.execute('ALTER TABLE customer ALTER COLUMN customer_state TYPE varchar(64)')

engine.execute('ALTER TABLE geolocation ALTER COLUMN geolocation_zip_code_prefix TYPE int
               USING geolocation_zip_code_prefix :: integer')
engine.execute('ALTER TABLE geolocation ALTER COLUMN geolocation_lat TYPE decimal')
engine.execute('ALTER TABLE geolocation ALTER COLUMN geolocation_lng TYPE decimal')
engine.execute('ALTER TABLE geolocation ALTER COLUMN geolocation_city TYPE varchar(64)')
engine.execute('ALTER TABLE geolocation ALTER COLUMN geolocation_state TYPE varchar(2)')

engine.execute('ALTER TABLE order_items ALTER COLUMN order_id TYPE varchar(64)')
engine.execute('ALTER TABLE order_items ALTER COLUMN product_id TYPE varchar(64)')
engine.execute('ALTER TABLE order_items ALTER COLUMN seller_id TYPE varchar(64)')
engine.execute('ALTER TABLE order_items ALTER COLUMN shipping_limit_date TYPE timestamp
               USING shipping_limit_date::timestamp without time zone')
engine.execute('ALTER TABLE order_items ALTER COLUMN price TYPE decimal')
engine.execute('ALTER TABLE order_items ALTER COLUMN freight_value TYPE decimal')

engine.execute('ALTER TABLE order_payments ALTER COLUMN order_id TYPE varchar(64)')
engine.execute('ALTER TABLE order_payments ALTER COLUMN payment_sequential TYPE int USING payment_sequential :: integer')
engine.execute('ALTER TABLE order_payments ALTER COLUMN payment_type TYPE varchar(20)')
engine.execute('ALTER TABLE order_payments ALTER COLUMN payment_installments TYPE int USING payment_installments :: integer')
engine.execute('ALTER TABLE order_payments ALTER COLUMN payment_value TYPE decimal')

engine.execute('ALTER TABLE reviews ALTER COLUMN order_id TYPE varchar(64)')
engine.execute('ALTER TABLE reviews ALTER COLUMN review_id TYPE varchar(64)')
engine.execute('ALTER TABLE reviews ALTER COLUMN review_score TYPE int USING review_score :: integer')
engine.execute('ALTER TABLE reviews ALTER COLUMN review_comment_title TYPE varchar(180)')
engine.execute('ALTER TABLE reviews ALTER COLUMN review_comment_message TYPE text')
engine.execute('ALTER TABLE reviews ALTER COLUMN review_creation_date TYPE timestamp
               USING review_creation_date::timestamp without time zone')
engine.execute('ALTER TABLE reviews ALTER COLUMN review_answer_timestamp TYPE timestamp
               USING review_answer_timestamp::timestamp without time zone')
```

Data Cleaning

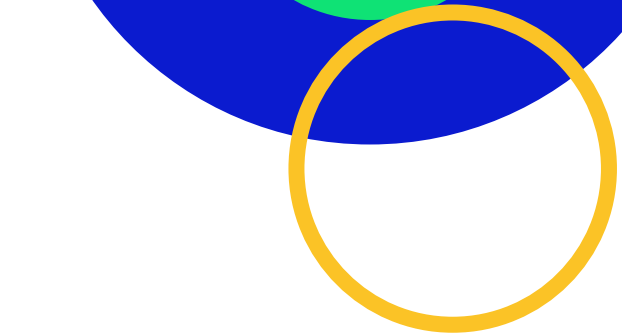
Cleaning done on Power BI

Product
-drop null values
-remove product name length, height, width, product category

Order
-split order_purchase 'timestamp' datatype to 'date' and 'time' columns

Review
-modify review_creation and review_timestamp to 'date' column
-remove duplicate review_id, order_id

Calendar
-add calendar
-only include 2015 onwards



Power BI

Problem Statement

Currently in Brazil e-commerce, there is an increase in big competitors such as Shopee penetrating into Brazil's Ecommerce market. How can Olist implement new strategies to protect our current market share and increase our year on year growth?

Business

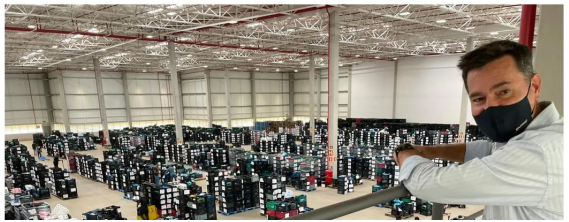
Shopee changes the game in Brazil's e-commerce sector

30 Aug 2021 05:38PM
(Updated: 30 Aug 2021 05

Brazilian social commerce marketplace Facily quietly raises $366M in less than a year, now valued at $850M

Mary Ann Azevedo @bayareawriter / 3:48 AM GMT+8 • November 17, 2021

Comment

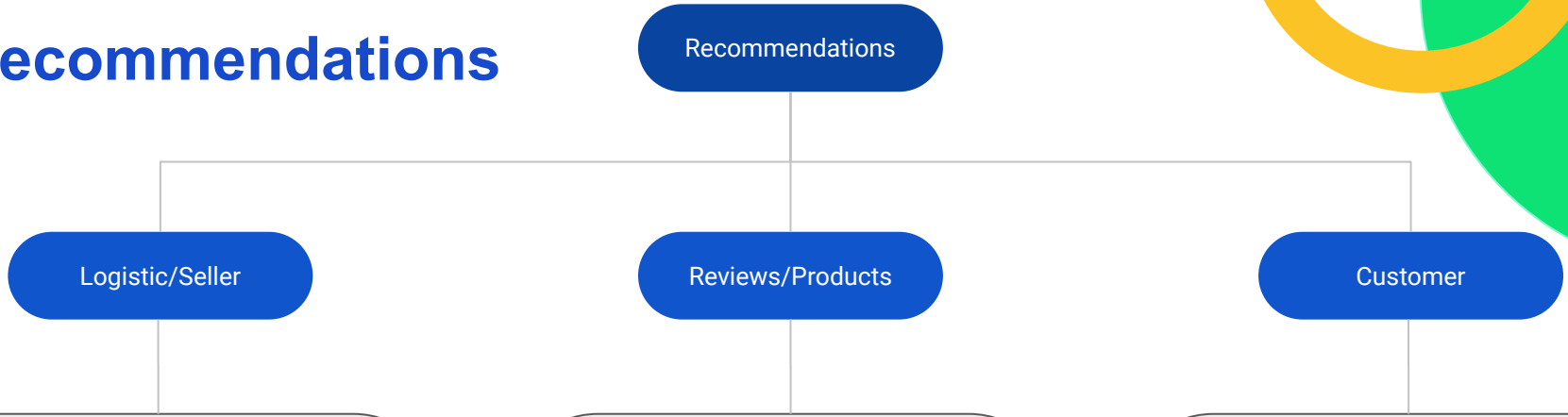


ECONOMY

Major E-commerce Players Have Stalled in Brazil—But the Fight Is Only Beginning

June 18, 2019

Recommendations



To deliver within estimated duration

R1: Improve inventory system to reduce delivery time
R2: Prevent extreme late deliveries by requiring sellers to update product availability in a timely manner
R3: Incentivise and penalise sellers in terms of commission taken by Olist accordingly to delivery performance

To change review system and incentivize sellers to improve review score

R1: Change review system - customer should not be able to review product without receiving item. Instead let customer to report this issue to CS, then have CS follow-up on the problem
R2: Incentivize sellers to hit high average review score consistently, by featuring them prominently/giving them free advertisements on Olist platform

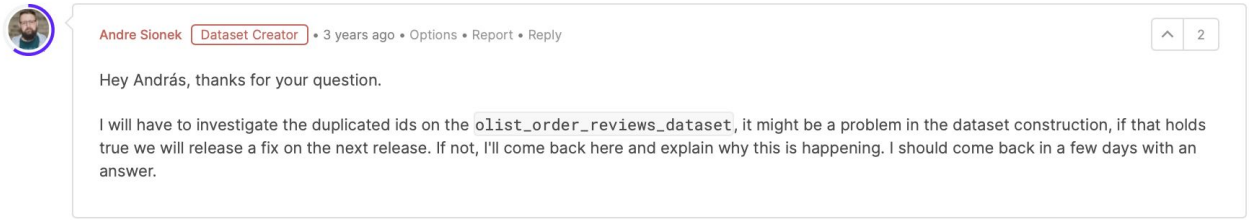
To improve customers lifetime value

R1: Based on customer categories (VVIP, Potential Customer, Loyal Customer, Big Spender, Lost Customer), recommend to deploy either one or combination of following strategies such as:

- notifying them of discount codes
- loyalty cards
- access to new products

Limitations of project

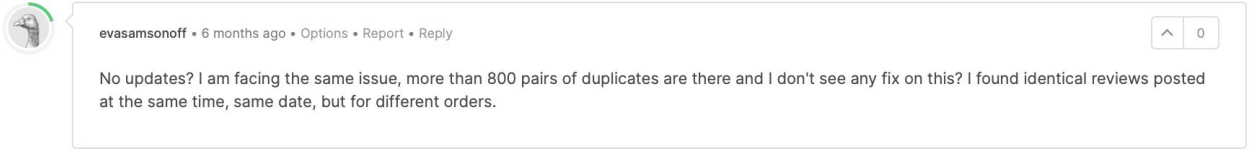
- Computer processor unable to handle translation
- Google API problem in translation
- Given dataset is from Sep 2016 to Aug 2018, but missing data from Nov-Dec 2016
- Data is not up to date
- Review dataset has problems



Andre Sionek Dataset Creator • 3 years ago • Options • Report • Reply 2

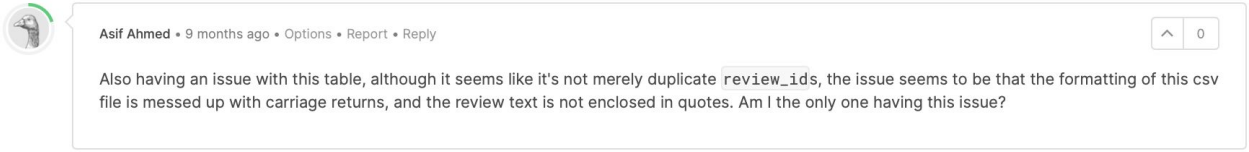
Hey András, thanks for your question.

I will have to investigate the duplicated ids on the `olist_order_reviews_dataset`, it might be a problem in the dataset construction, if that holds true we will release a fix on the next release. If not, I'll come back here and explain why this is happening. I should come back in a few days with an answer.



evasamsonoff • 6 months ago • Options • Report • Reply 0

No updates? I am facing the same issue, more than 800 pairs of duplicates are there and I don't see any fix on this? I found identical reviews posted at the same time, same date, but for different orders.



Asif Ahmed • 9 months ago • Options • Report • Reply 0

Also having an issue with this table, although it seems like it's not merely duplicate `review_id`s, the issue seems to be that the formatting of this csv file is messed up with carriage returns, and the review text is not enclosed in quotes. Am I the only one having this issue?

Retrieved from:
https://www.kaggle.com/olistbr/brazilian-ecommerce/discussion/71650

Q&A