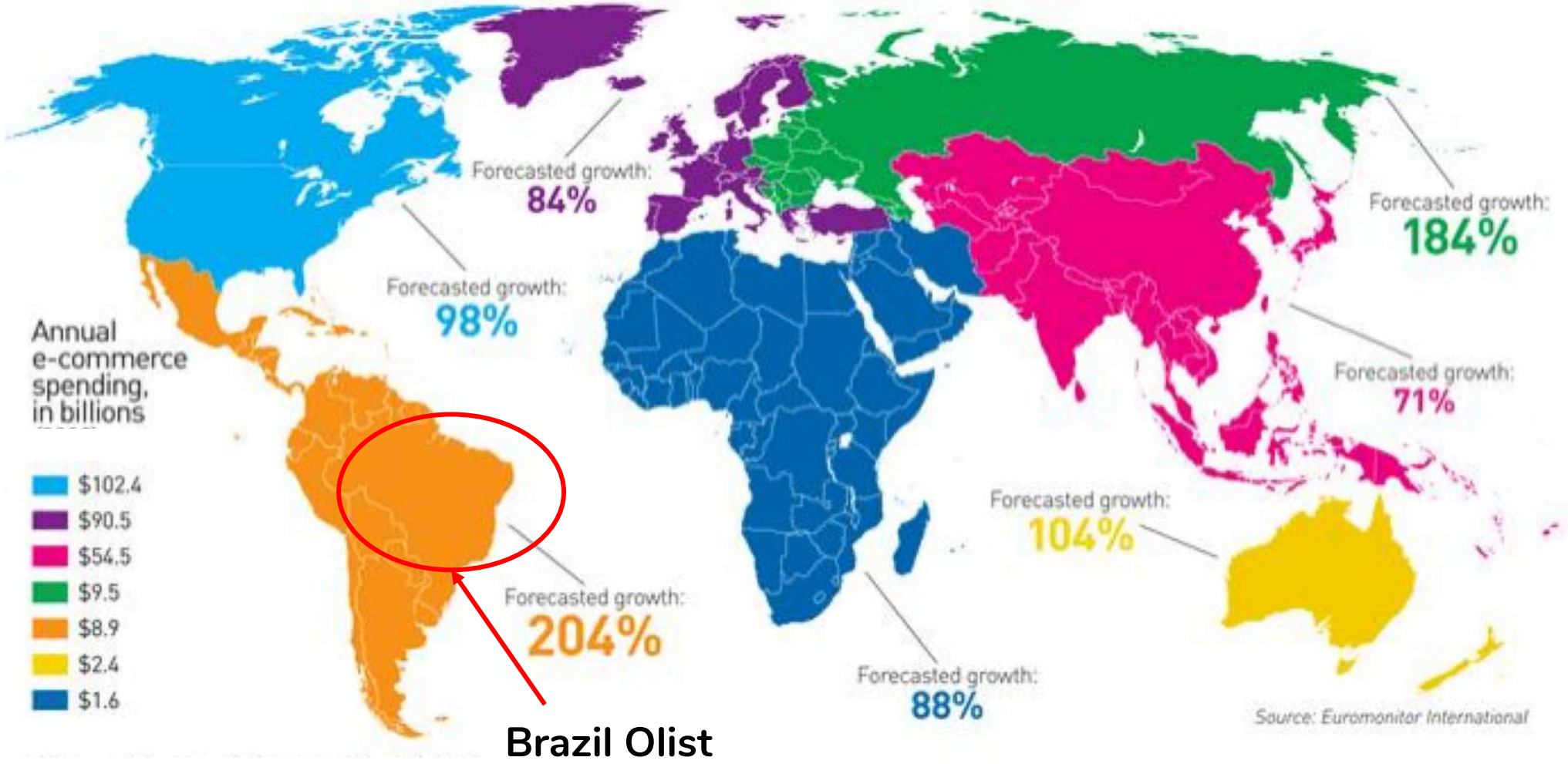
Big Data Analytics in E-commerce Platform

Motivation



#Kaggle Brazilian Olist Dataset

Agenda

Data Aggregation & Visualization

- Data Schema
- Feature Engineering

Sales Analysis:

- Based on Sales Volume
- Based on Product Categories
- Based on Sellers
- Based on Payment Method

NLP

- Sentimental Analysis
- Effect on Ratings

Clustering

- Total Payment
- Freight/Ratio

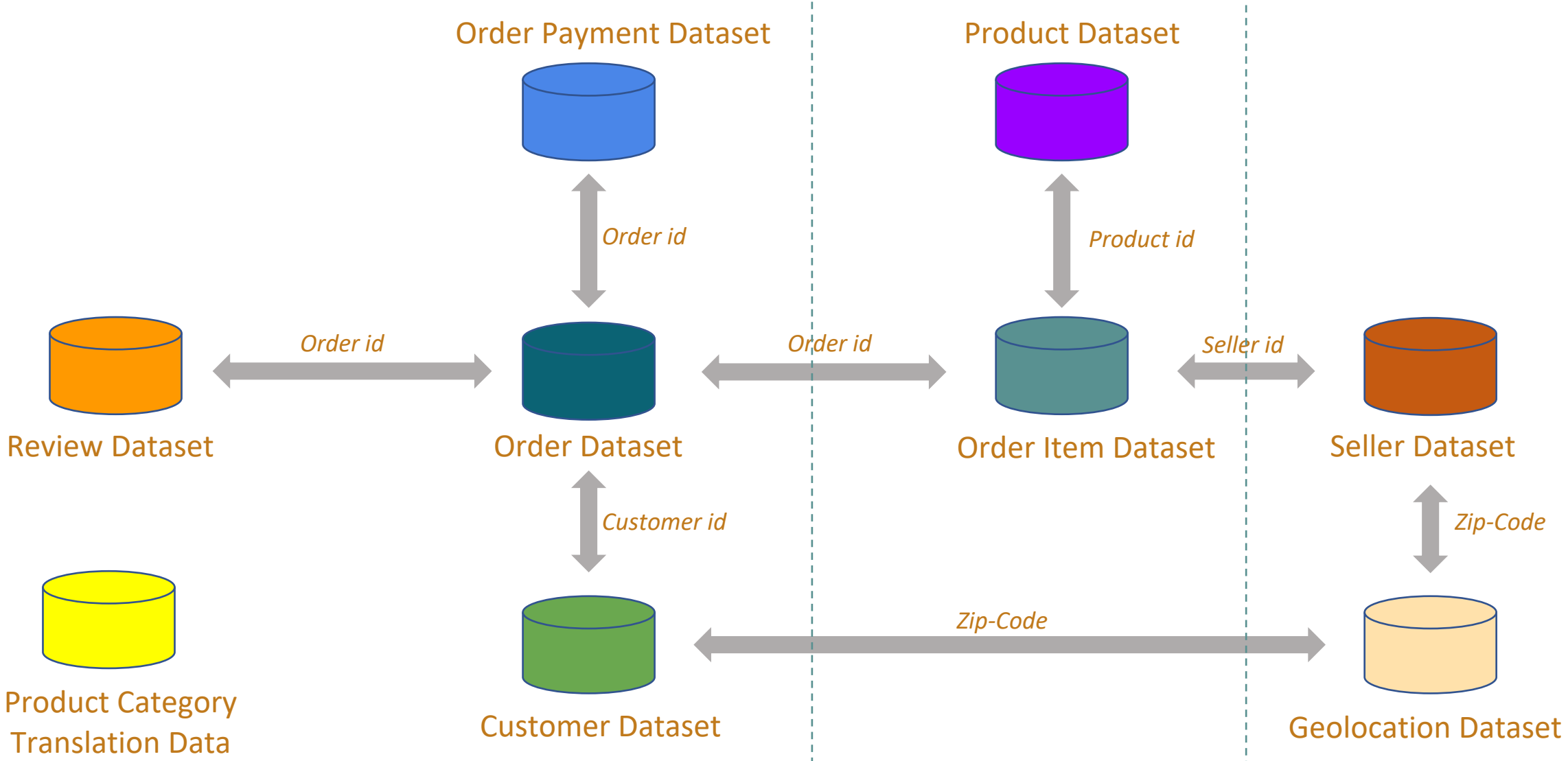
Classification

- Binary (Late/On-Time)
- Multiple (Ratings)

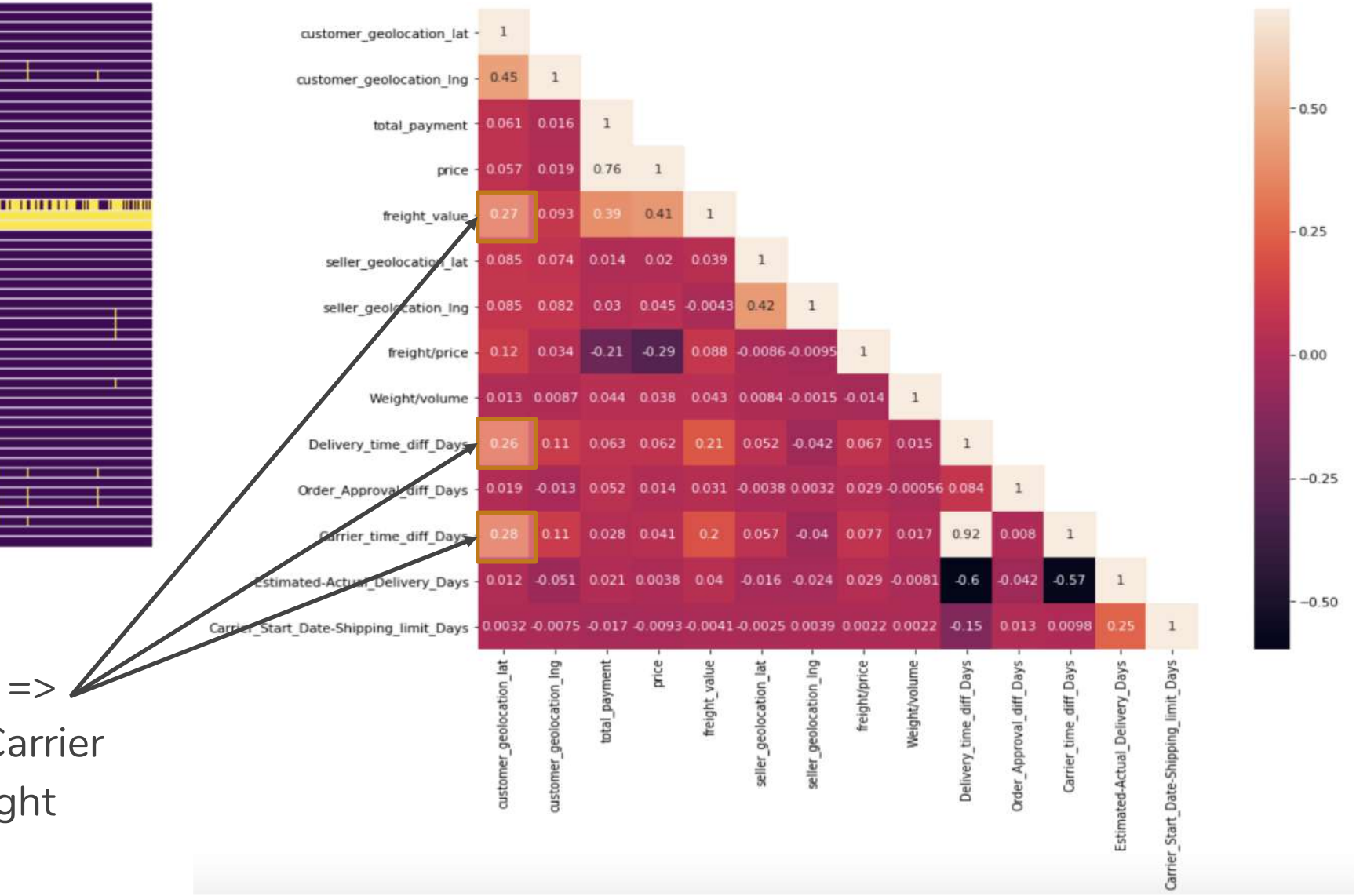
Consolidation and Recommendations

Data Schema

Brazilian Olist Dataset



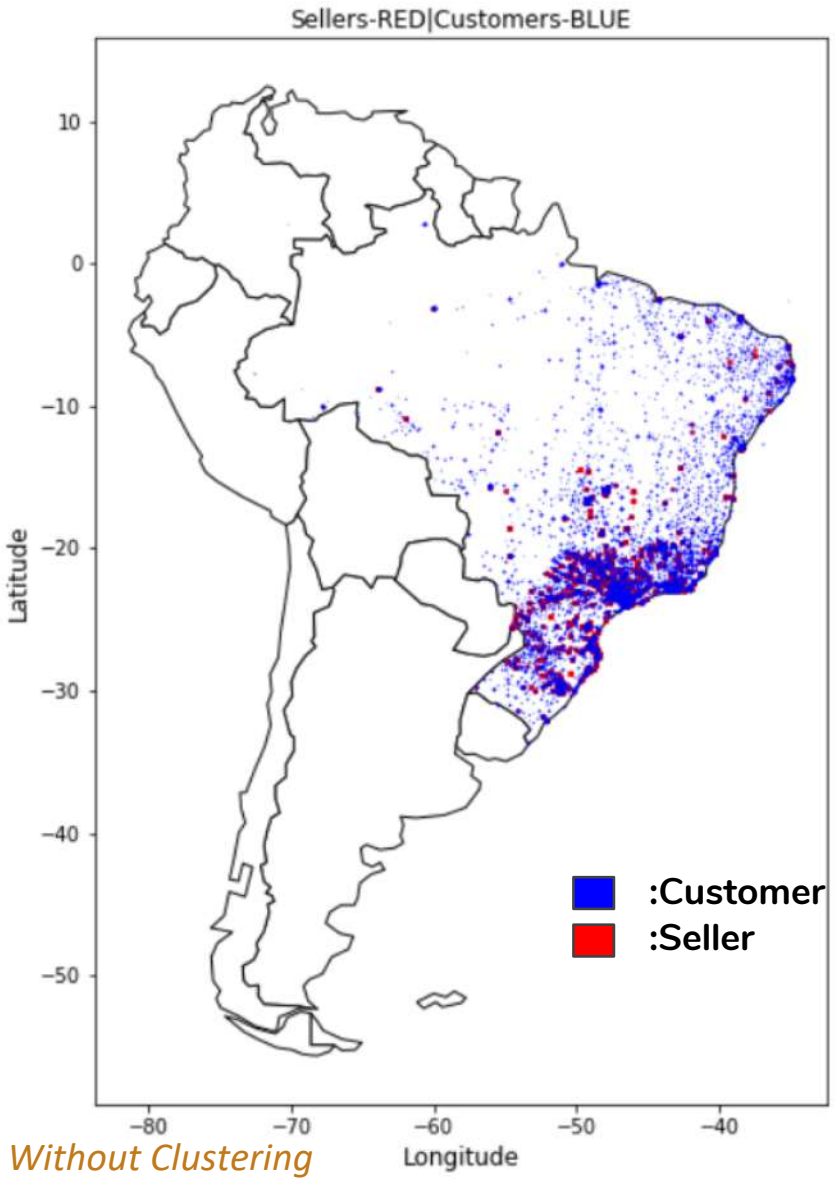
Feature Engineering



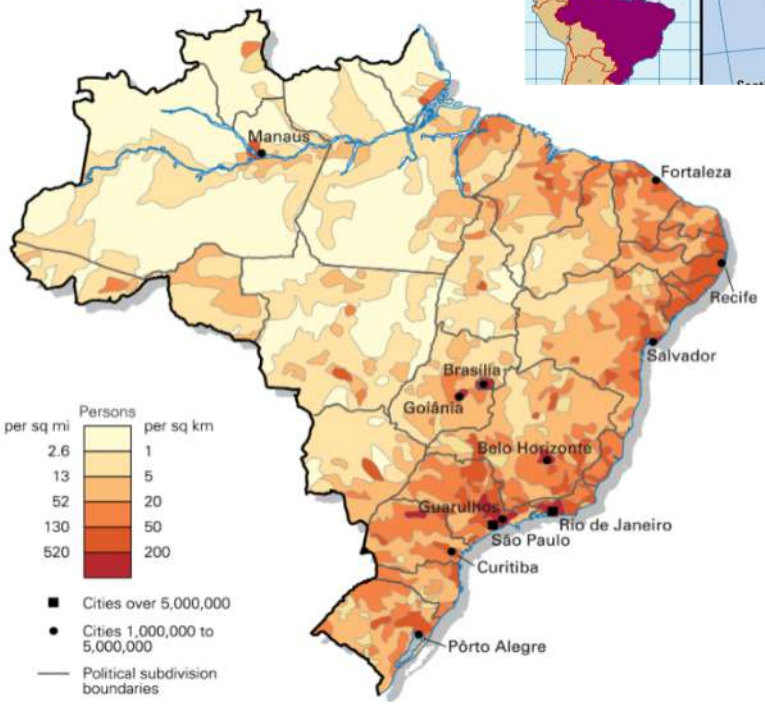
Correlations:

- Customer geolocation Lat => Positive correlation with Carrier Time, Delivery time & Freight Values

Seller & Customer Visualization



Without Clustering



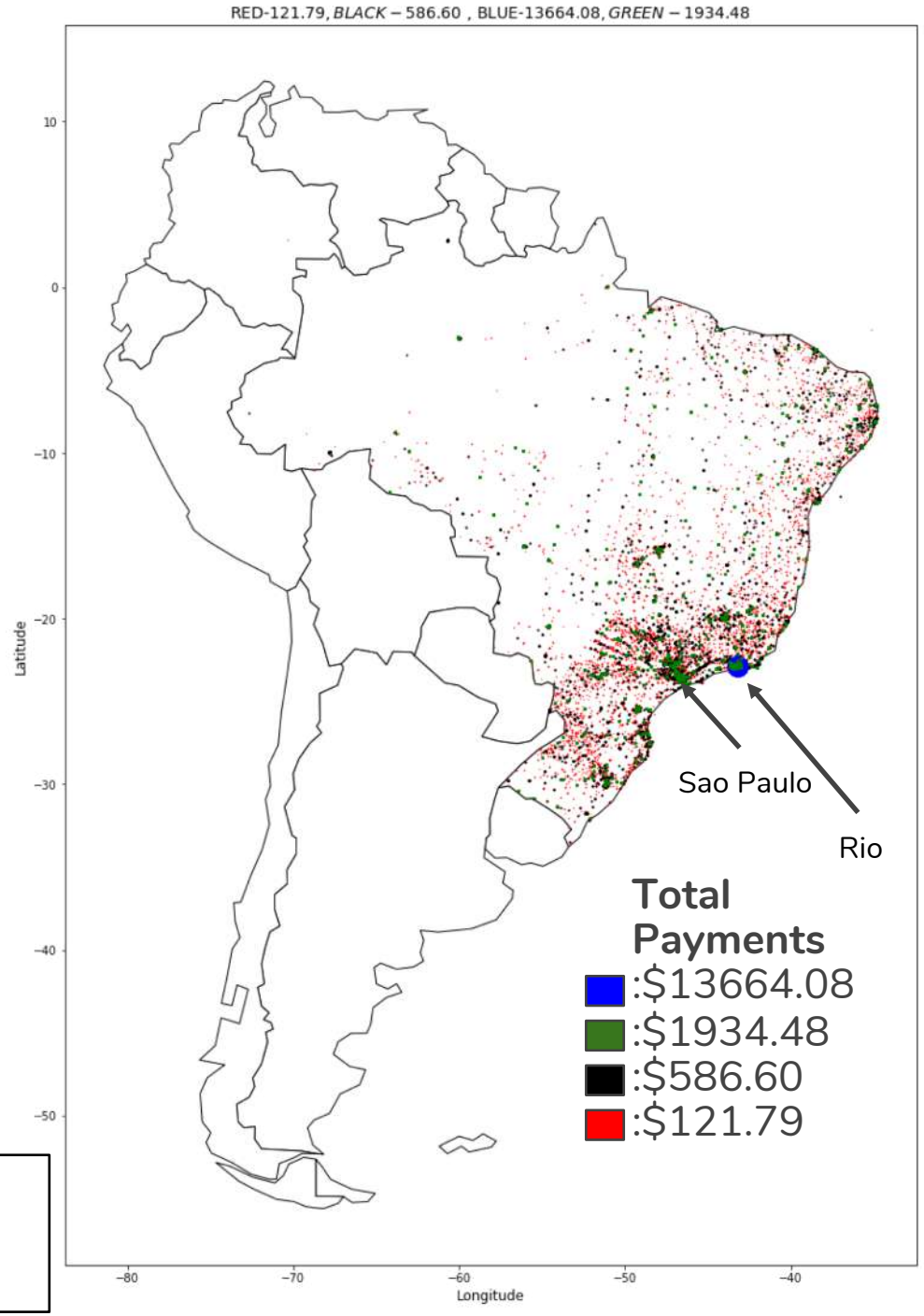
2018 rank	City	State	2018 Estimate
1	*São Paulo*	São Paulo	12,176,866
2	*Rio de Janeiro*	Rio de Janeiro	6,688,927
3	*Brasília*	Distrito Federal	2,974,703
4	*Salvador*	Bahia	2,857,329
5	*Fortaleza*	Ceará	2,643,247
6	*Belo Horizonte*	Minas Gerais	2,501,576
7	*Manaus*	Amazonas	2,145,444
8	*Curitiba*	Paraná	1,917,185
9	*Recife*	Pernambuco	1,637,834
10	*Goiânia*	Goiás	1,495,705

https://www.britannica.com/place/Brazil

Clustering

Total Payments(mean)

- Most of the revenue came from South and the Southeast regions of Brazil.
- It is also possible to see that large cities and capitals, where population is bigger, have larger participation on revenue.



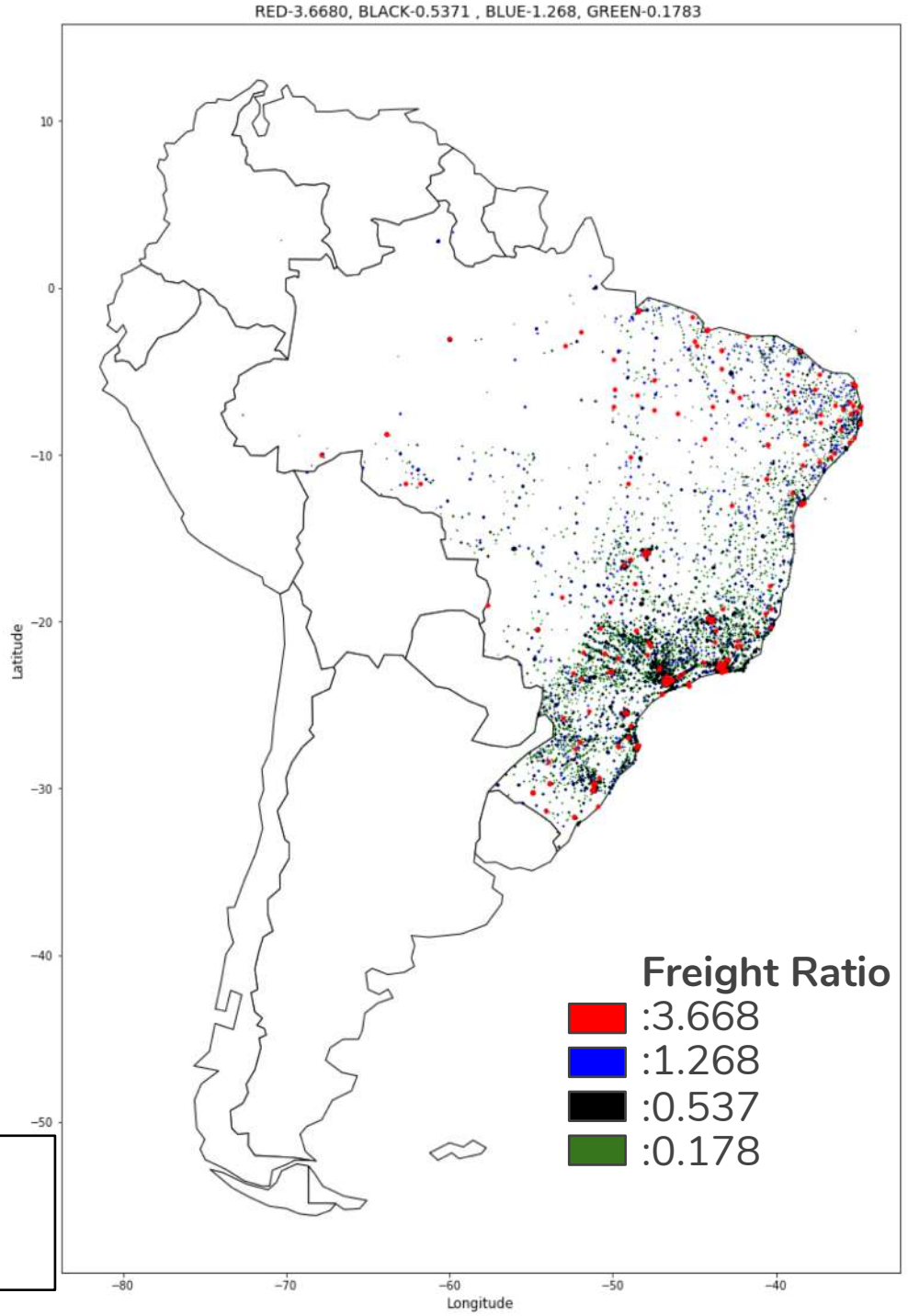
Method: Map/Reduce, K-Means Clustering, Geo-Pandas

Clustering

- ## Freight Ratio

  This ratio indicates the percentage of the product price that a person had to pay just to get their order delivered.
  For Example:
  Price=$100 , Freight=$20.
  Ratio indicates 20% goes to Transportation.

  - Due to logistics costs, we see lower freight ratios in densely populated areas and higher freight ratios on sparsely populated regions.
  - The Higher Ratio in Densely Populated regions might be due to the **Express** delivery.

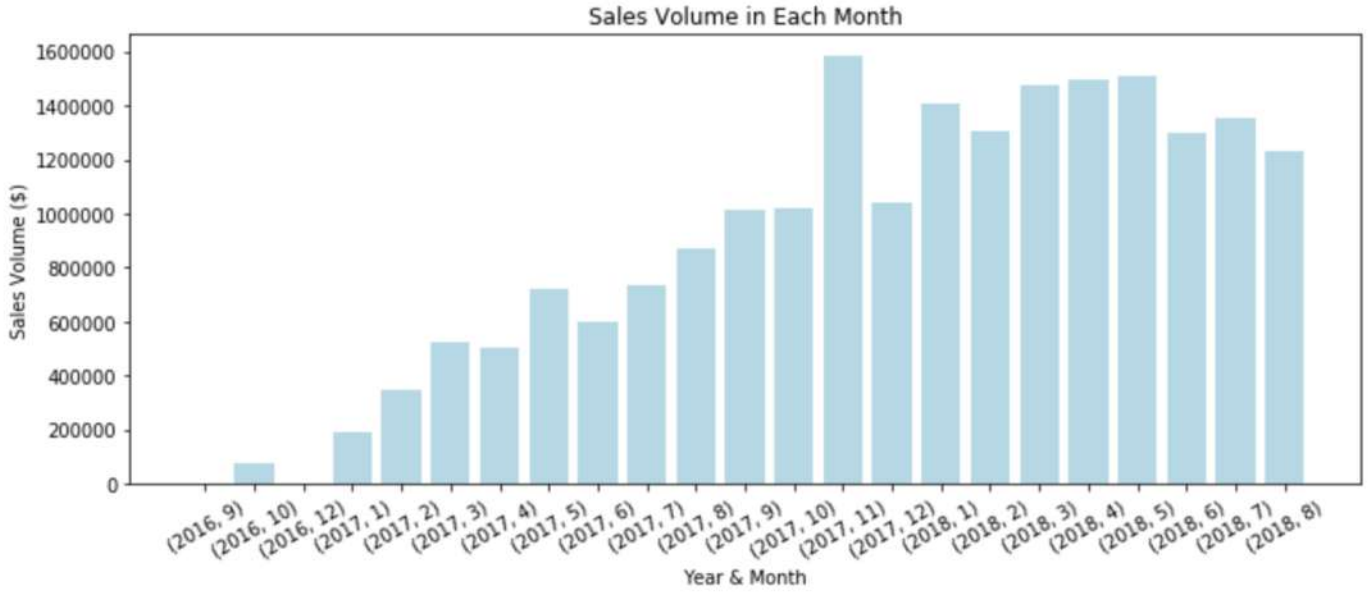


Method: Map/Reduce, K-Means Clustering, Geo-Pandas

Sales Analysis

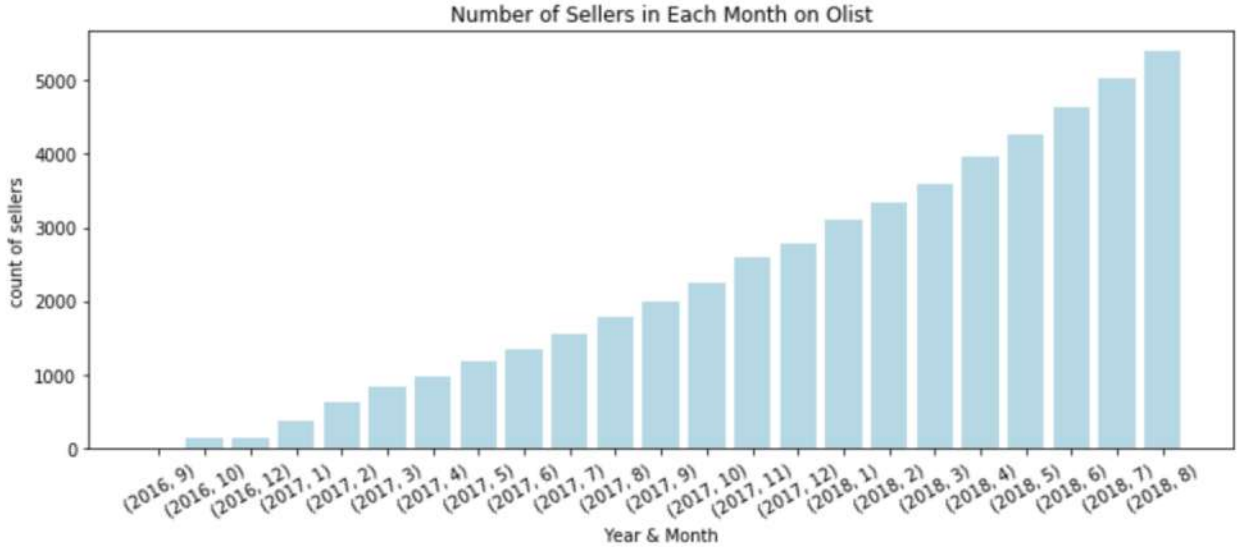
Based on Sales Volume

- Sales volume increased before Oct, 2017, while went steady after that.



Based on Sellers

- Number of sellers increased steadily during 2016-2018.



Method: Map/Reduce, frequent item

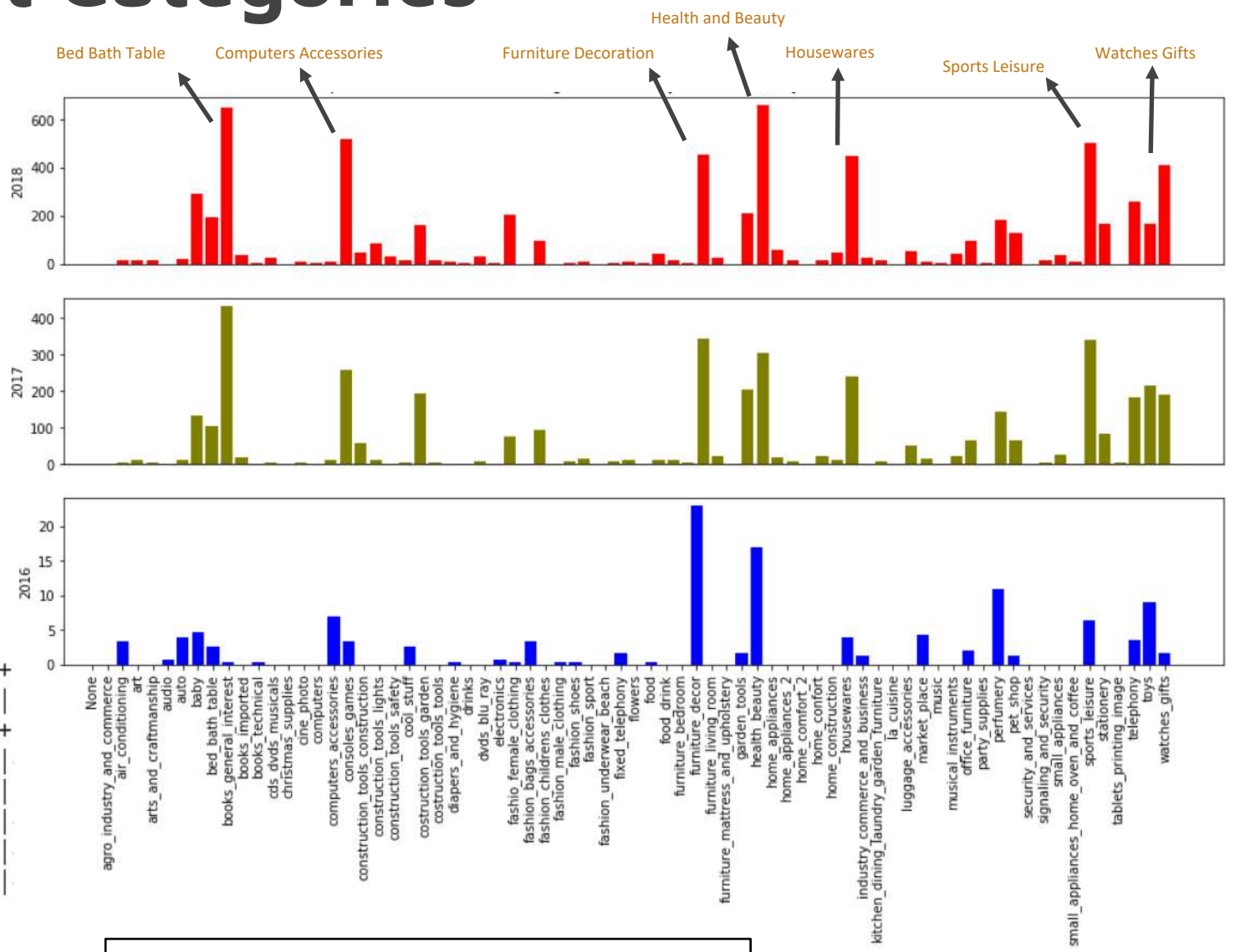
The Data for the month of Nov 2016 and Dec 2016 had very less Records

Based on Product Categories

Most popular categories:
- Health & Beauty
- Furniture Decoration
- Bed-Bath-Table

Categories often bought together (Market basket analysis)
- Home comfort & Bed-Bath-Table

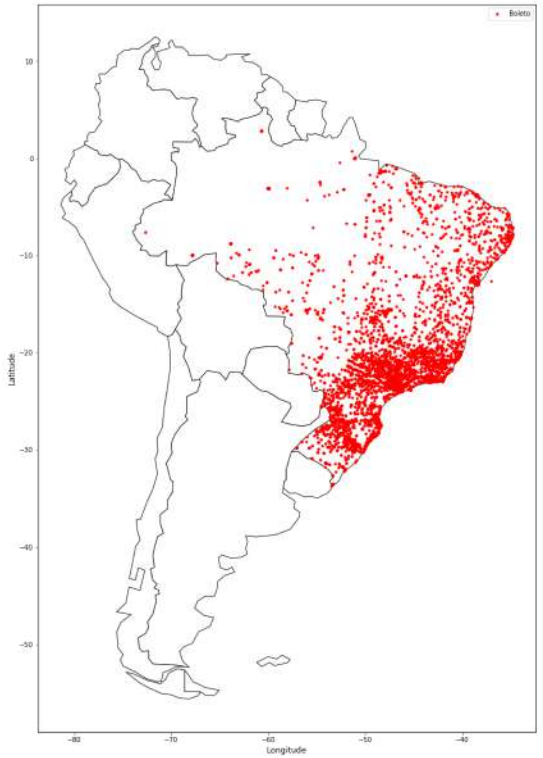
product_id_1	product_id_2	frequency
3f14d740544f37ece...	36f60d45225e60c7d...	12
98d61056e0568ba04...	060cb19345d90064d...	6
5b8a5a9417210b1b8...	e5ae72c62ebfa7086...	6
5fc3e6a4b52b0c414...	5d790355cbeded0cd...	6
36f60d45225e60c7d...	e53e557d5a159f5aa...	34



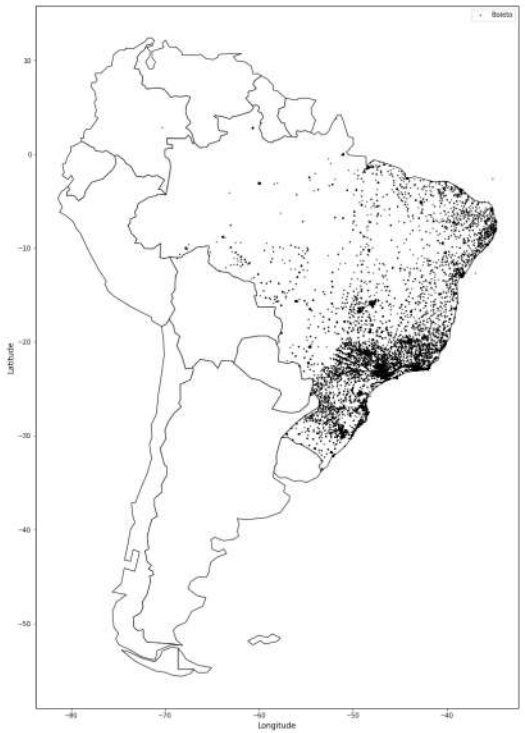
Method: Map/Reduce, frequent item

The Data for the month of Nov 2016 and Dec 2016 had very less Records

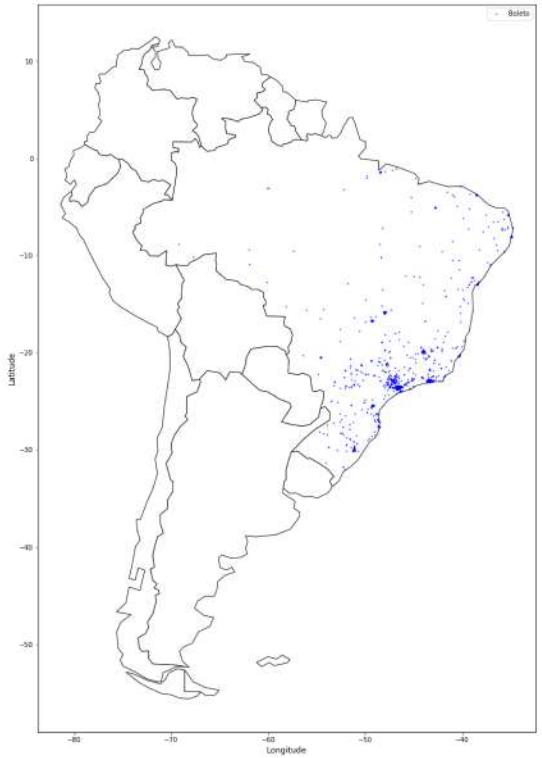
Dominant Payment Method



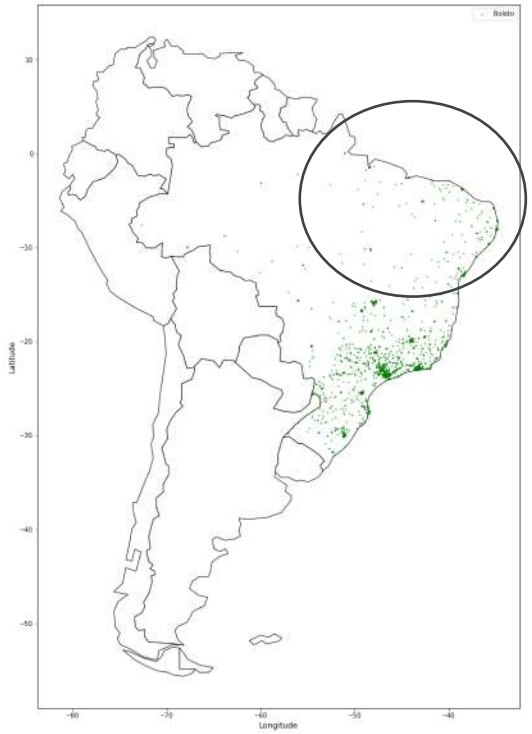
Boleto (Check)



Credit Card



Debit Card



Voucher (Coupon)

```
Dominant payment method :Boleto = 18905
Dominant payment method :Credit Card = 71654
Dominant payment method :Debit Card = 1497
Dominant payment method :Voucher = 3364
```

Method: Map/Reduce, K-Means Clustering, Geo-Pandas

Binary classification: On-time delivery

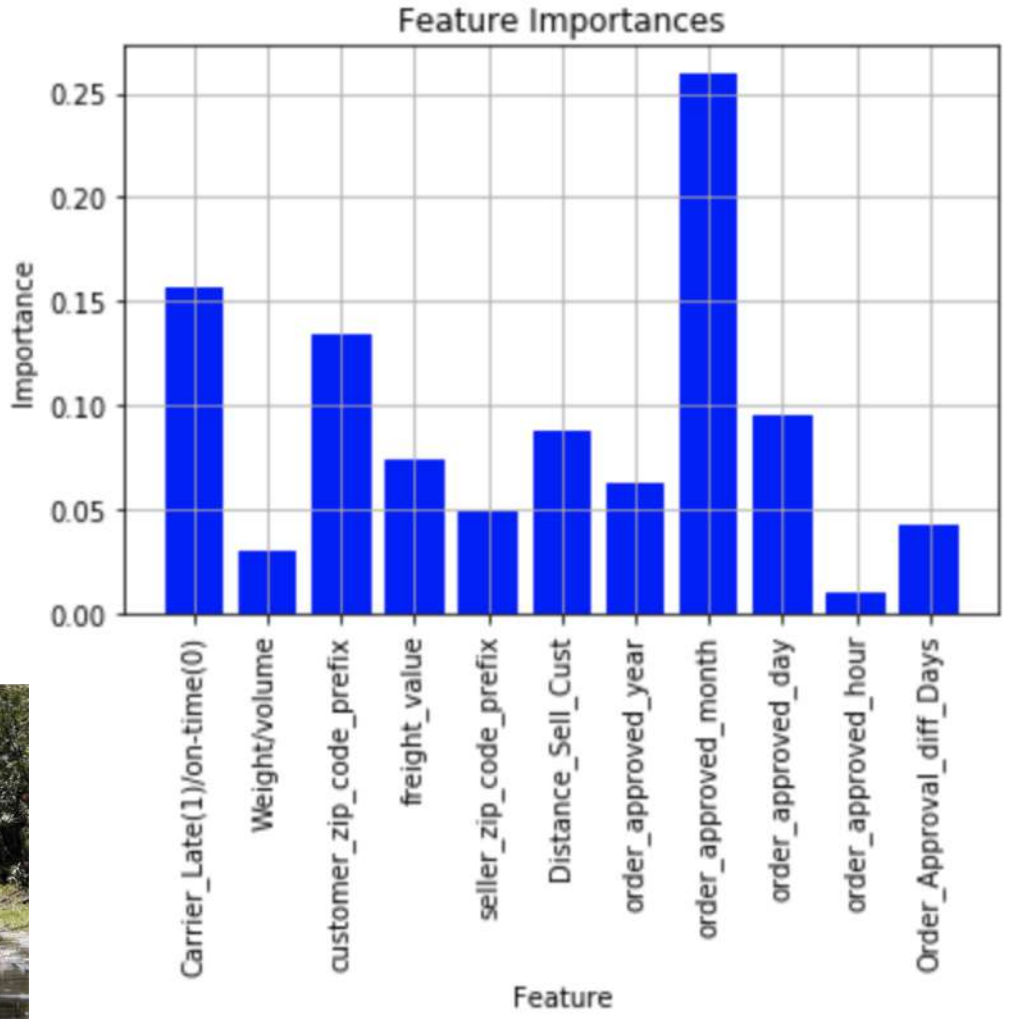
Binary classification models: Predict the on-time delivery with binary classification models

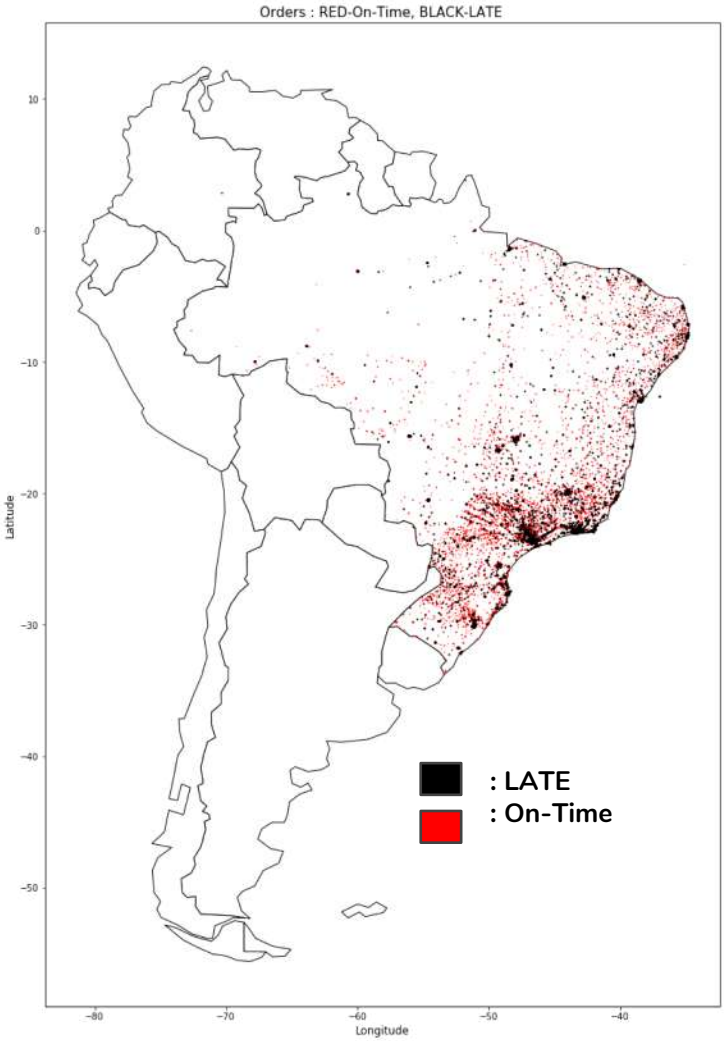
- **Input variables:** 'Carrier_Late/on-time', 'Weight/volume', 'Customer_zip_code', 'Freight_value', 'Seller_zip_code', 'Distance_seller_customer', 'Order_approved_year/month/day/hour' ,'Order_approval_diffDays'
- **Target variables:** 'On-time delivery'
- **ML models:** Logistic regression, Decision tree (Random forest, Gradient Boosting tree), Naive Bayes, Linear SVM, Neural Network
- **Best model:** Gradient Boosting tree (80% accuracy)
- **Hyperparameter Tuning:**
  - Identifying optimal parameters
  - handling imbalanced datasets(classWeights)

Binary classification: On-time delivery

Feature importances:

- Customer ZIP code
- Carrier Delay
- Order Approval Month





Multiclass classification: Review scores

Multi-class classification models: Predict the review scores with multi-class classification models

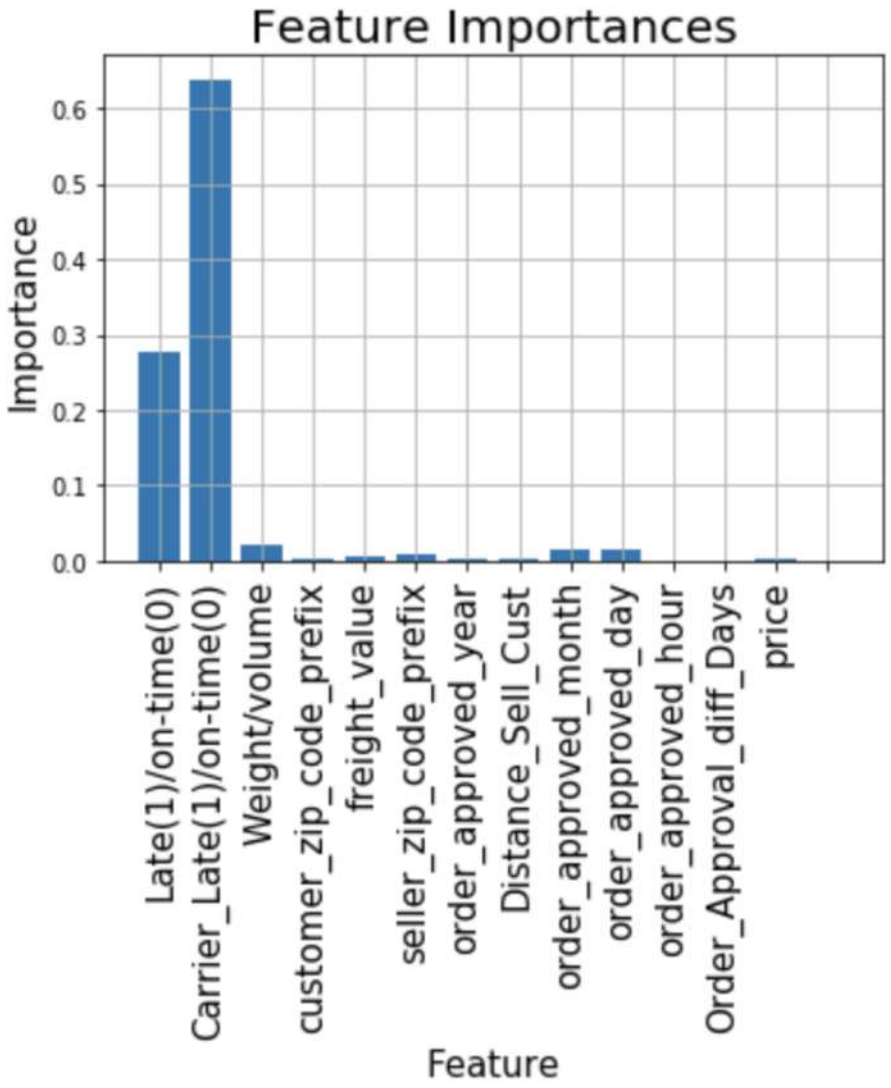
- **Input variables:** 'Late(1)/on-time(0)','Carrier_Late(1)/on-time(0)','weight/volume','customer_zip_code','freight_value', 'product_height_cm', 'product_length_cm', 'product_weight_g','product_width_cm', 'seller_zip_code','distance_seller_customer', 'order_approved_year/month/day/hour/,''Order_Approval_diff_Days','price'
- **Target variables:** '**review_score**'
- **ML models:** Logistic regression, Decision tree (Random forest), Naive Bayes, Neural Network
- **Best model:** Neural Network
- **Hyperparameter Tuning:** Identifying optimal parameters

Multiclass classification: Review scores

Feature importances:

- On-time delivery :

Customers value Delivery on-time more.



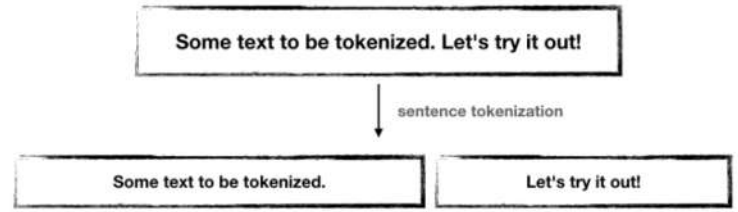
Natural language Processing

Step 1

Review Comment Messages (all lower case)

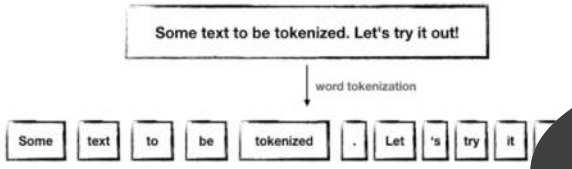
Step 2

Sentence tokenization



Step 3

Word tokenization



Step 4

Stopwords, Punctuation

```
Any time of the day is a good time for pie
```

```
Any time day good time pie
```

Step 5

Stemming and Lemmatization

adjustable → adjust
formality → formaliti
formaliti → formal

was → (to) be
better → good
meeting → meeting

Step 6

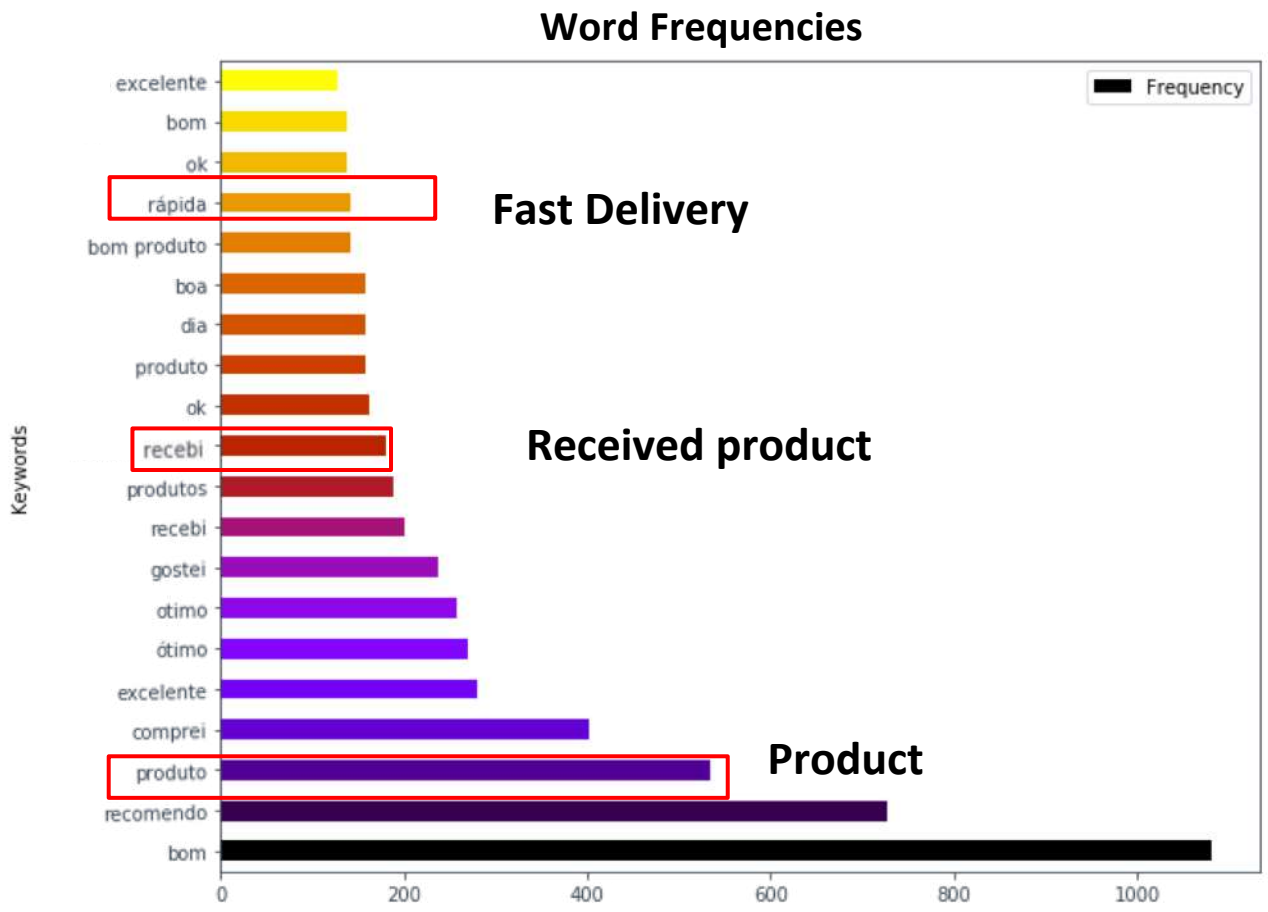
Chunking, Chinking and POS tagging

Results

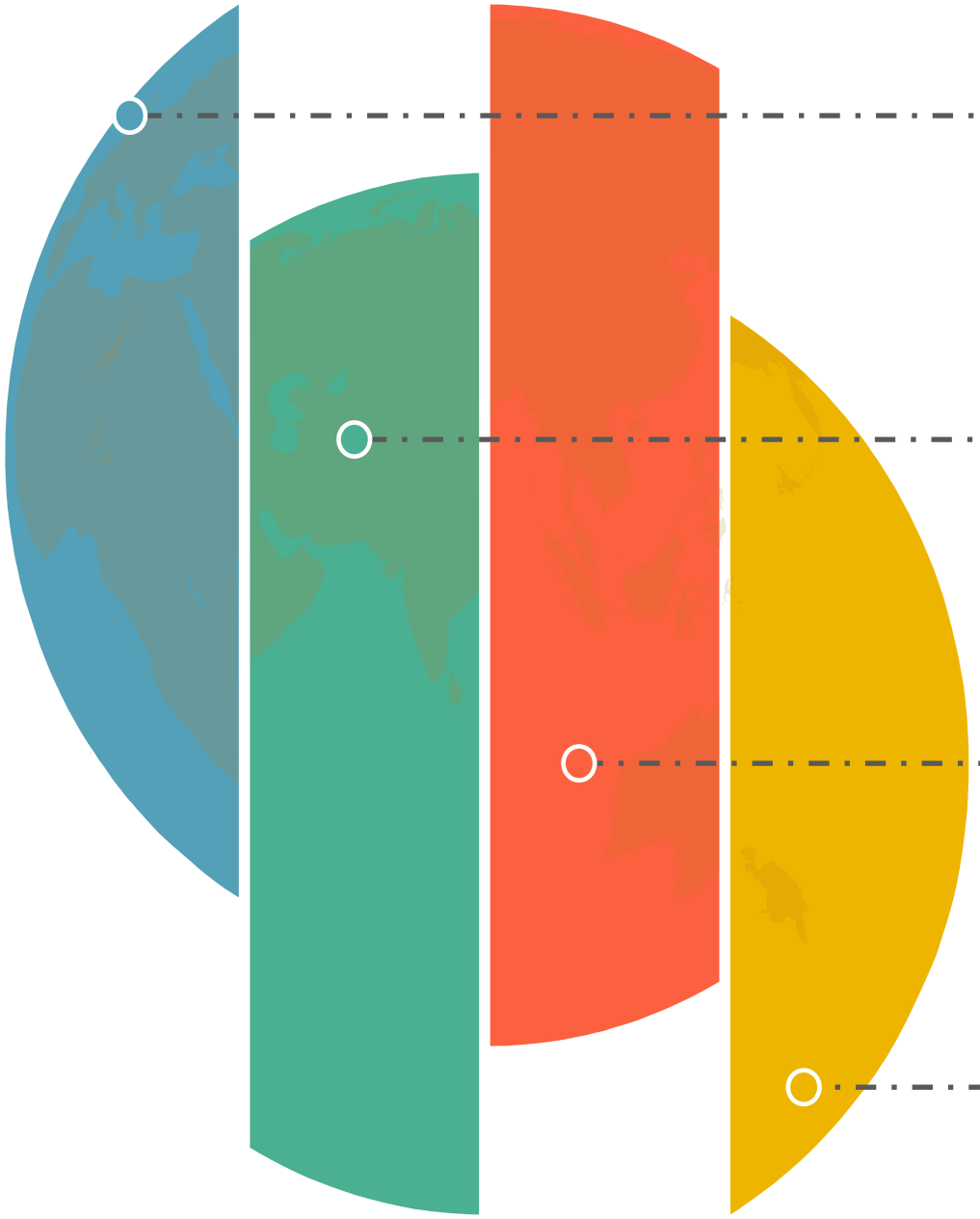
Natural language Processing

Let's extract the top 20 keywords and Visualizing the output

excellent
Good product
Everything is ok
Fast
Good
Good
Day
Great product
Okay
Receive
Products
Received
Like
Great
Great
Excellent
Purchase
Product
Recommend
Good



Challenges



A. Data

- Massive amounts of data, Joining datasets, Missing values , Intuitive trends.
- Null values, different dates,different sources, parsing the data

B. ARIMA Model

- Low time series Data Points

C. Only ID

- No Product name
- No actual Market Basket Analysis possible.

D. Portuguese

- Comments in different language for NLP : Sentimental Analysis

Recommendations

01 **Sellers**

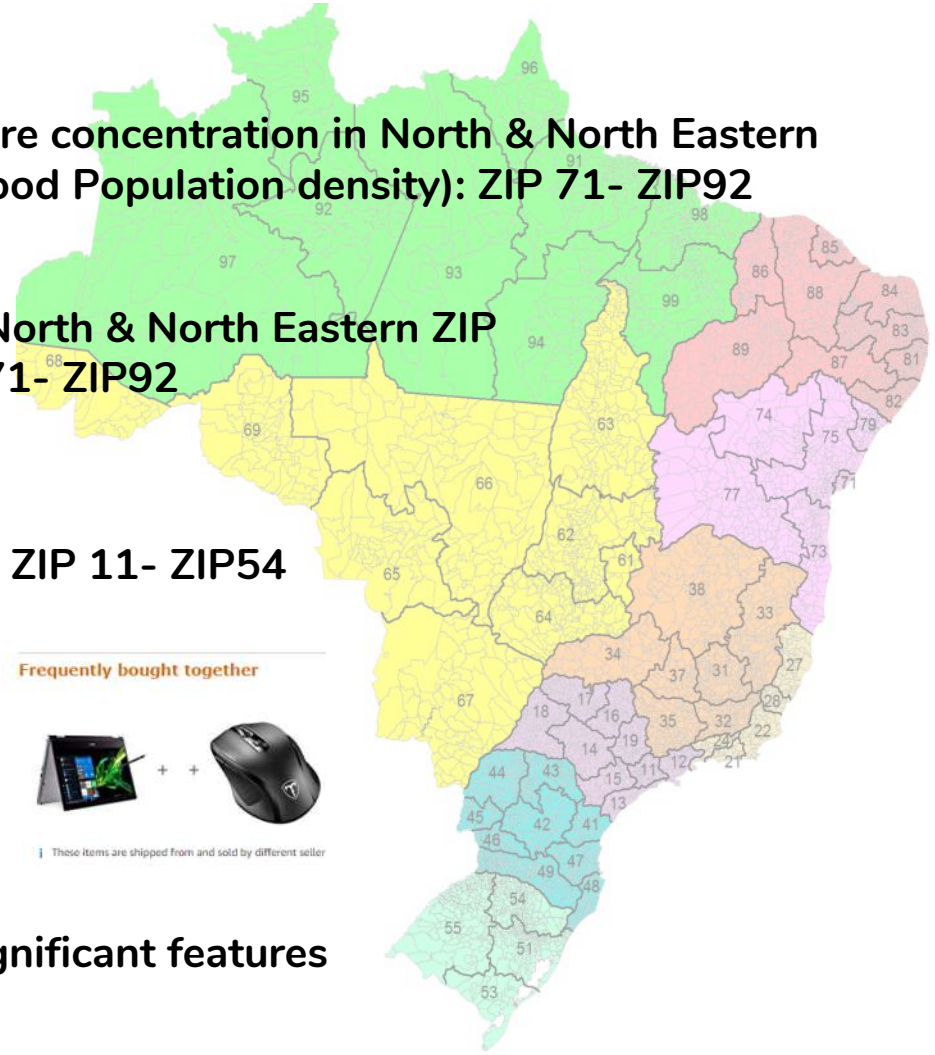
02 **Targeted Marketing Coupons**

03 **Carrier Delay**

04 **Frequently Bought Together**

05 **On-Time Delivery**

More concentration in North & North Eastern (Good Population density): ZIP 71- ZIP92



North & North Eastern ZIP 71- ZIP92

ZIP 11- ZIP54

Most significant features

#WikiPedia

рахмат
danke
謝謝
ngiyabonga
شكراً جزيلاً
Баярлалаа
spas
mersi
kia ora
barka
welalin
tack
teşekkür ederim
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
kiitos
dankie
thank you
gracias
хвала
nanni
nandri
dhanyavad
mauruuru
köszönöm
ačiū
asante
manana
obrigada
tenki
bedankt
bayarlalaa
gràçie
hvala
dankon
akun
djiere dieuf
tau
mochchakkeram
murakoze
enkosi
dziękuję
дякую
mamnun
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
chokrane
sobodi
dēkuji
sagolun
sukriya
kop khun krap
taiku
obrigado
mèsi
didi madloba
ありがとう
arigatô
takk
dakujem
trugarez
najis tuke
tanemirt
grazie
terima kasih
rahmet
εὐχαριστώ
dhanyavadagalu
shukriya
mercê
мерси
kam sah hamnida
rahmat
diolch
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci