olist

Brazilian E-Commerce - FY19 marketing insight

CX1115 mini-project

Liang Xuchao
Lin Yan
Liao Zixin

Scenario

Olist.com is preparing for the coming FY19 marketing strategy plan. ***Olist*** invited NTU students as group of Data Analyst Consultants to look for marketing insights in order to improve Sales amount and revenue, based on the 2016-2018 ***Olist*** customer and order data.

- Which factors will affect Sales Amount and Sales Revenue?
- Hypothesis: review score, delivery time, and proof
  - Linear Regression
  - Random Forest Regression
- FY19 Marketing plan: Proposal to improve Sales amount and revenue

Data Preparation

Data Preparation

Data source

- Brazilian E-Commerce Public Dataset By ***Olist***

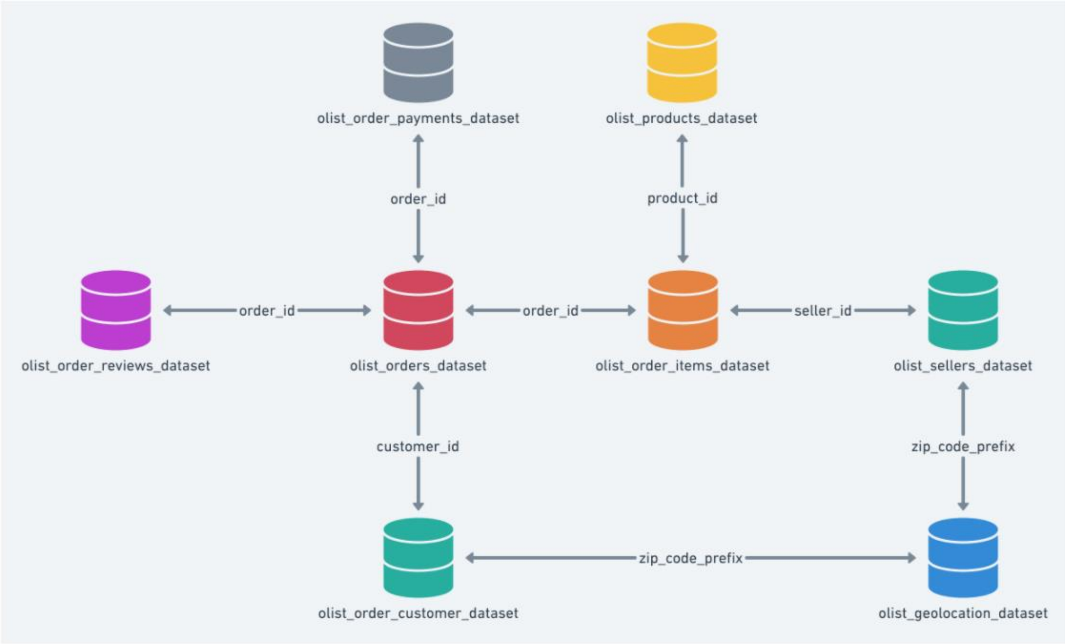
  https://www.kaggle.com/olistbr/brazilian-ecommerce

Data Preparation - Call Api for Data

1. Using Kaggle api feature to download data through command line
2. Unzip and extract files in the current folder

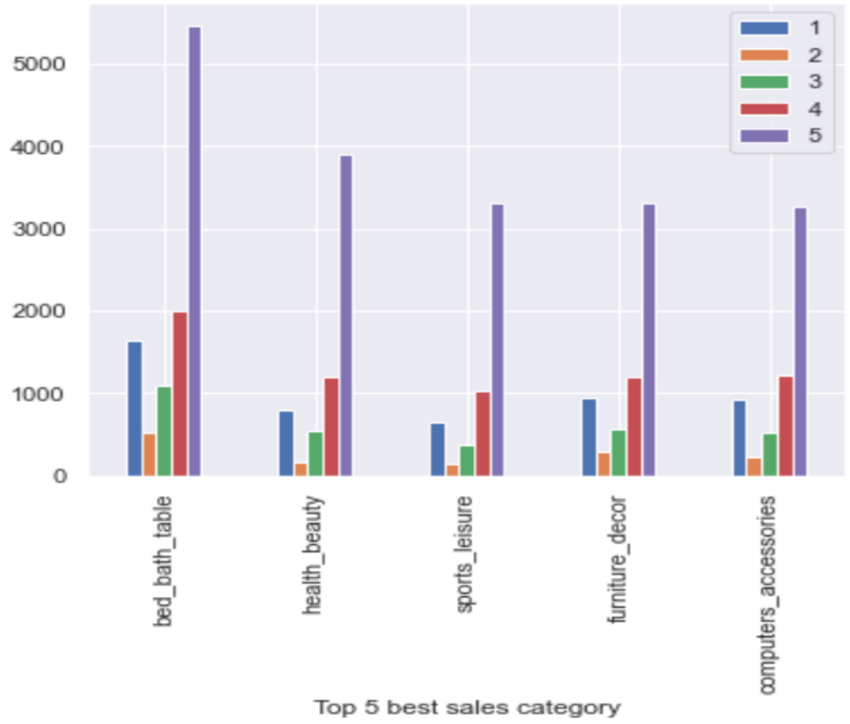
```
import os
from zipfile import ZipFile

# app.py
# call api
# 1. According to kaggle website, api call require install package in pip -> pip install )
# 2. After install go to kaggle account page  https://www.kaggle.com/<username>/account to get api token
# 3. Place downloaded json file C:\Users\<Windows-username>\.kaggle\kaggle.json
os.system("kaggle datasets download -d olistbr/brazilian-ecommerce")
with ZipFile('brazilian-ecommerce.zip', 'r') as zipObj:
   # Extract all the contents of zip file in current directory
   zipObj.extractall()


os.system("kaggle datasets download -d olistbr/marketing-funnel-olist")
with ZipFile('marketing-funnel-olist.zip', 'r') as zipObj:
   # Extract all the contents of zip file in current directory
   zipObj.extractall()
```

Data Preparation - Data mapping



Factors Analysis

Analysis - Sales Amount vs Average Review Score (Linear Regression)



Evaluate the impact of review score on sales.

Analysis - Sales Amount vs Average Review Score (Linear Regression)

Sales Amount of seller for each category:

- Group by 'seller_id' & 'product_category_name_english' and drop the sales order whose statue is 'canceled' & ‘unavailable’.
- Return the sum of sales order of each seller for each category.

Average review score of seller for each category

- Group by 'seller_id' & 'product_category_name_english' and drop the sales order whose statue is 'canceled' & ‘unavailable’.
- Return the mine of review score of each seller for each category.

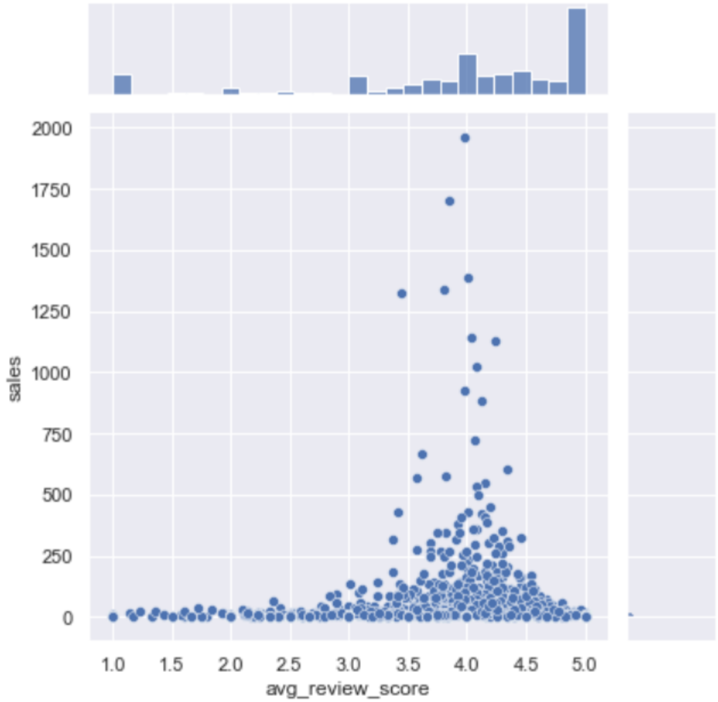
Analysis - Sales Amount vs Average Review Score (Linear Regression)

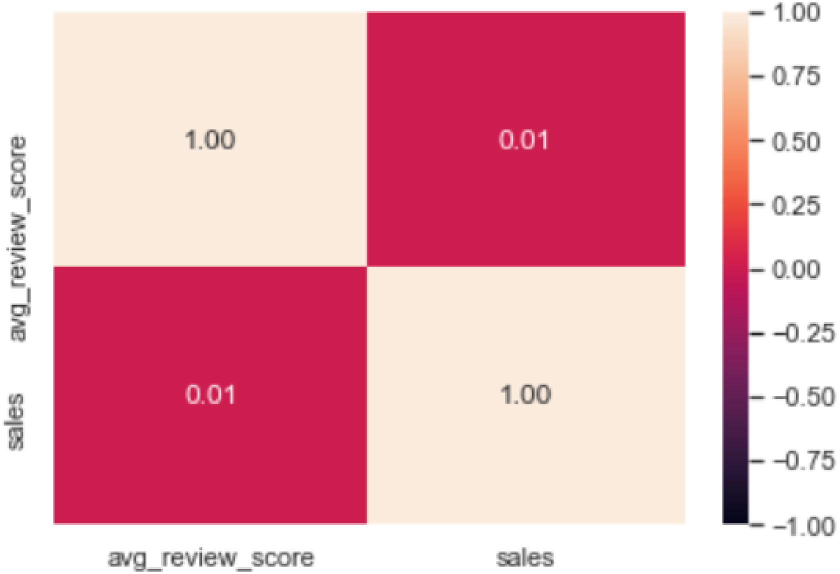
	seller_id	product_category_name_english	avg_review_score	sales
3	289cdb325fb7e7f891c38608bf9e0962	perfumery	4.577586	116.0
5	66922902710d126a0e7d26b0e3805106	pet_shop	4.441718	163.0
6	2c9e548be18521d1c43cde1c582c6de8	stationery	3.755556	135.0
7	8581055ce74af1daba164fdbd55a40de	auto	4.231441	458.0
9	16090f2ca825584b5a147ab24aa30c86	auto	4.050584	257.0
11	7c67e1448b00f6e969d365cea6b010ab	office_furniture	3.439909	1323.0
12	7c67e1448b00f6e969d365cea6b010ab	office_furniture	3.439909	1323.0
13	001cca7ae9ae17fb1caed9dfb1094831	garden_tools	3.864486	214.0
14	001cca7ae9ae17fb1caed9dfb1094831	garden_tools	3.864486	214.0
15	87142160b41353c4e5fca2360caf6f92	computers_accessories	4.349481	289.0
17	1900267e848ceeba8fa32d80c1a5f5a8	bed_bath_table	3.821366	571.0
21	ea8482cd71df3c1969d7b9473ff13abc	telephony	3.939095	1215.0
22	d2374cbcbb3ca4ab1086534108cc3ab7	bed_bath_table	3.610526	665.0
23	70a12e78e608ac31179aea7f8422044b	telephony	3.653846	78.0
24	70a12e78e608ac31179aea7f8422044b	telephony	3.653846	78.0
25	70a12e78e608ac31179aea7f8422044b	telephony	3.653846	78.0
27	cc419e0650a3c5ba77189a1882b7556a	health_beauty	4.032371	1143.0
28	8b321bb669392f5163d04c59e235e066	electronics	3.980498	923.0

The Sellers who received the review amount more than 50.

Analysis - Sales Amount vs Average Review Score (Linear Regression)

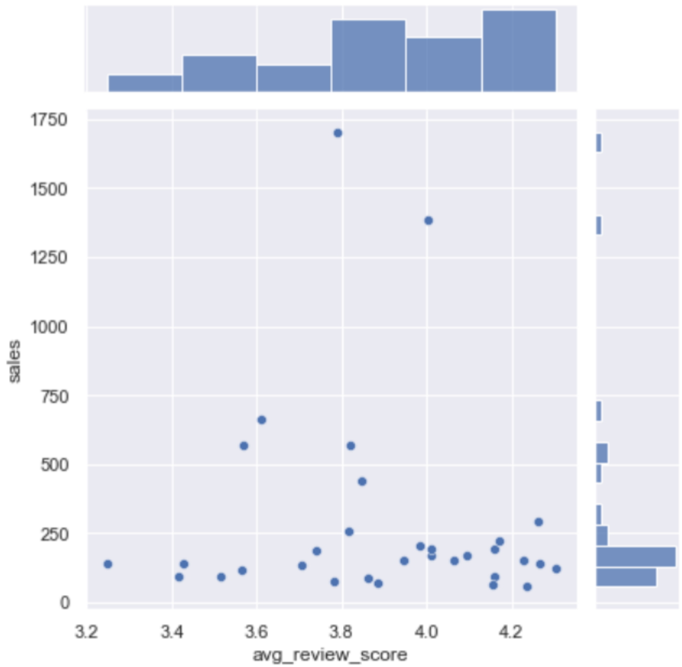
Data Set: the sales from all category

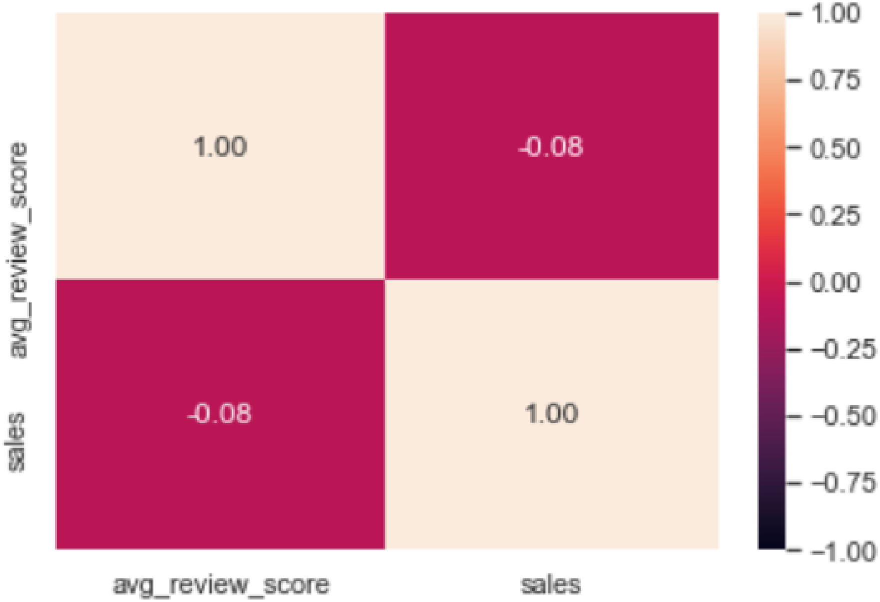




Analysis - Sales Amount vs Average Review Score (Linear Regression)

Data Set: the sales for category ‘Bed Bath Table’





Analysis - Sales Amount vs Average Review Score (Linear Regression)

Statistical intuition --------- No dependence

Correlation of all : 0.01

Correlation of category ‘Bed Bath Table : -0.08

Analysis - Random Forest Regression for multiple Predictors

Data filtering & processing:

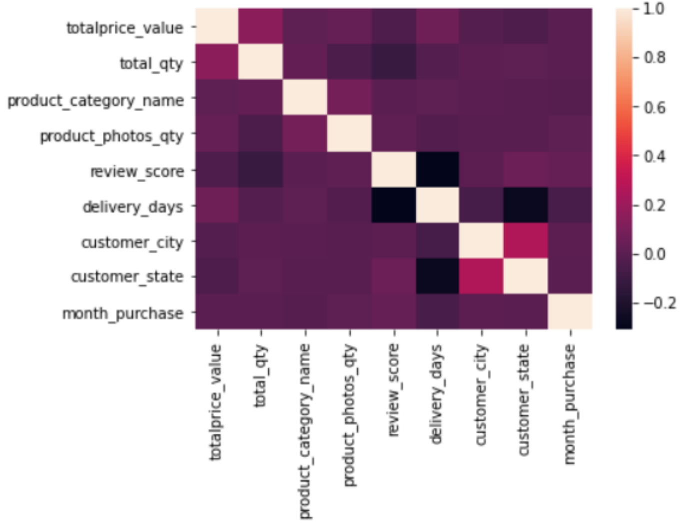
- Drop rows with many missing values
- Convert string to numeric value using LabelEncoder() for regression

Predictors included in regression:

1. Product photo quantity
2. Review
3. Product Category
4. Delivery time
5. Customer City

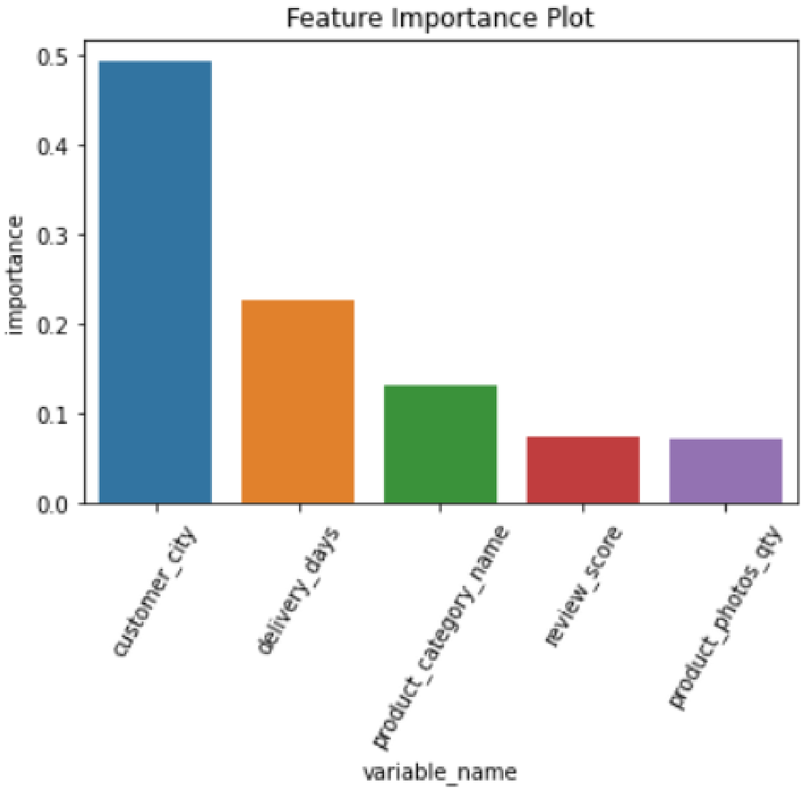
Analysis - Random Forest Regression for multiple factors

```
Training Features Shape: (72705, 6)
Testing Features Shape: (24235, 6)
Training labels Shape: (72705, 1)
Testing labels Shape: (24235, 1)
Mean Absolute error in predicting train data 0.1027016477072858
Mean Absolute error in predicting test data 0.26625151016348647
```



Train Set: 75%
Test Set: 25%

Analysis - Random Forest Regression for multiple factors



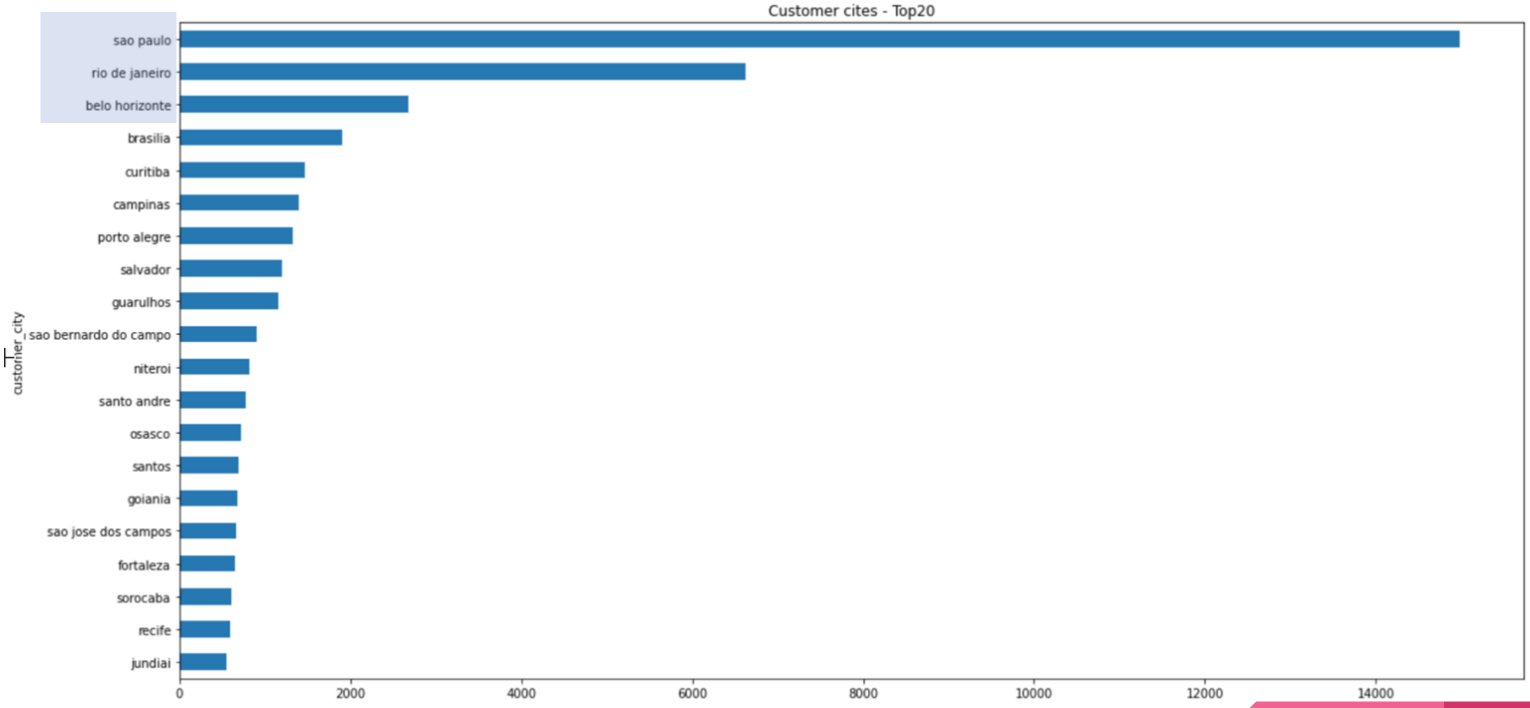
Results for review score matched linear regression result

The top 3 factors added up have around 90% of importance on predict sales amount

Proposal - FY19 Marketing plan

Factor 1: Customer City

Where did the customer come from?



Factor 2: Delivery Time

```
[173]: fast_seller['delivery_time'].describe()

[173]: count    110831.000000
       mean          9.143741
       std           8.638963
       min         -16.000000
       25%           4.000000
       50%           7.000000
       75%          12.000000
       max         205.000000
       Name: delivery_time, dtype: float64
```

Proposal - ***Olist*** to promote the fast sellers in top 3 cities

'sao paulo' fast seller list

```
[163]: sp = pd.DataFrame(fast_seller[fast_seller['sell
       sp.groupby(['seller_id'])
       sp = pd.DataFrame(sp[sp['delivery_time'] <= 0])
       sp.count()

[163]: seller_id                  129
       delivery_time              129
       seller_zip_code_prefix     129
       seller_city                129
       seller_state               129
       dtype: int64

[179]: print(sp['seller_id'])

       5680      ea8482cd71df3c1969d7b9473ff13abc
       9257      8b321bb669392f5163d04c59e235e066
       9351      8b321bb669392f5163d04c59e235e066
       9630      8b321bb669392f5163d04c59e235e066
```

** fast definition: delivery time = 0 day

Proposal - ***Olist*** to promote the fast sellers in top 3 cities

'rio de janeiro' fast seller list:

```
[164]: rdj = pd.DataFrame(fast_seller[fast_seller['seller_city'] == 'rio de janeiro'])
       bh.groupby(['seller_id'])
       rdj = pd.DataFrame(rdj[rdj['delivery_time'] <= 0])
       rdj.count()
```

```
[164]: seller_id                12
       delivery_time            12
       seller_zip_code_prefix   12
       seller_city              12
       seller_state             12
       dtype: int64
```

```
[176]: print(rdj['seller_id'])
```

```
20806     f84a00e60c73a49e7e851c9bdca3a5bb
27786     7a425d299613df3e613bcf9d2eaf5c49
40018     46dc3b2cc0980fb8ec44634e21d2718e
40133     46dc3b2cc0980fb8ec44634e21d2718e
40464     46dc3b2cc0980fb8ec44634e21d2718e
97246     db46ca7bce82b11f7e247539271fc390
```

** fast definition: delivery time = 0 day

Proposal - ***Olist*** to promote the fast sellers in top 3 cities

'belo horizonte' fast seller list:

[172]:
```
bh = pd.DataFrame(fast_seller[fast_seller['seller
bh = pd.DataFrame(bh[bh['delivery_time'] <= 0])
bh.count()
```

[172]:
```
seller_id                11
delivery_time            11
seller_zip_code_prefix   11
seller_city              11
seller_state             11
dtype: int64
```

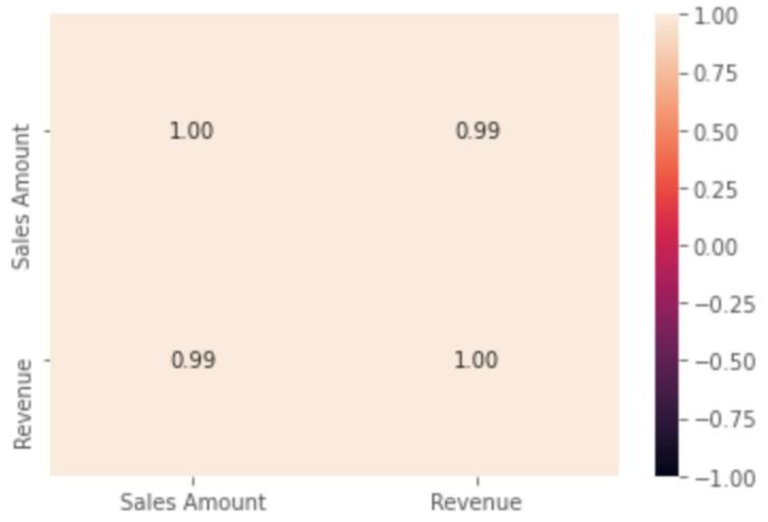
[175]:
```
print(bh['seller_id'])
```

```
14408    85d9eb9ddc5d00ca9336a2219c97bb13
14461    85d9eb9ddc5d00ca9336a2219c97bb13
14855    85d9eb9ddc5d00ca9336a2219c97bb13
66448    fc906263ca5083d09dce42fe02247800
70674    282f23a9769b2690c5dda22e316f9941
```

** fast definition: delivery time = 0 day

Factor 3: Sales Amount and Sales Revenue regarding Product Category

[351]: <AxesSubplot:>



Conclusion: Sales Amount and Sales Revenue have **strong** relationship.

The more Sales Amount, the more Sales Revenue.

Analysis - Top 10 Sales Revenue and Sale Amount regarding product category

- Top 10 Sales **Revenue** product:

	product_category_name_english	Revenue
0	health_beauty	1302046.97
1	watches_gifts	1254322.95
2	bed_bath_table	1107397.98
3	sports_leisure	1029631.88
4	computers_accessories	950134.59
5	furniture_decor	772496.16
6	housewares	668880.94
7	cool_stuff	664637.13
8	auto	618395.50
9	garden_tools	519473.33

- Top 10 Sales **Amount** product:

	product_category_name_english	Sales Amount
0	bed_bath_table	11990
1	health_beauty	10033
2	sports_leisure	9005
3	furniture_decor	8833
4	computers_accessories	8151
5	housewares	7380
6	watches_gifts	6213
7	telephony	4726
8	garden_tools	4590
9	auto	4400

Analysis - product category with good selling performance

Top Line Product Category Hierarchy 1 :

`health_beauty`, `watches_gifts`,

`bed_bath_table`, `sports_leisure`

rank	product_category_na	Revenue	Sales Amount
1	health_beauty	1302046.97	10033
2	watches_gifts	1254322.95	6213
3	bed_bath_table	1107397.98	11990
4	sports_leisure	1029631.88	9005
5	computers_accessories	950134.59	8151
6	furniture_decor	772496.16	8833
7	housewares	668880.94	7380
8	cool_stuff	664637.13	3999
9	auto	618395.5	4400
10	garden_tools	519473.33	4590

Good potential product:

`watches_gifts`

Analysis - product category with good selling performance

Top Line Product Category Hierarchy 2 :

`computers_accessories`, `furniture_decor`, `housewares`

rank	product_category_na	Revenue	Sales Amount
1	health_beauty	1302046.97	10033
2	watches_gifts	1254322.95	6213
3	bed_bath_table	1107397.98	11990
4	sports_leisure	1029631.88	9005
5	computers_accessories	950134.59	8151
6	furniture_decor	772496.16	8833
7	housewares	668880.94	7380
8	cool_stuff	664637.13	3999
9	auto	618395.5	4400
10	garden_tools	519473.33	4590

Good potential product:

`computers_accessories`

Analysis - product category with good selling performance

Top Line Product Category Hierarchy 3 :

`cool_stuff`, `auto`, `garden_tools`

rank	product_category_na	Revenue	Sales Amount
1	health_beauty	1302046.97	10033
2	watches_gifts	1254322.95	6213
3	bed_bath_table	1107397.98	11990
4	sports_leisure	1029631.88	9005
5	computers_accessories	950134.59	8151
6	furniture_decor	772496.16	8833
7	housewares	668880.94	7380
8	cool_stuff	664637.13	3999
9	auto	618395.5	4400
10	garden_tools	519473.33	4590

Good potential product:

`cool_stuff`

Analysis: Seasonal Product

`computers` in Q3

`toys` in Q4,

	product_category_name_english	Q1 Sales Revenue
0	health_beauty	324729.40
1	computers_accessories	322707.61
2	sports_leisure	322322.77
3	watches_gifts	294718.51
4	bed_bath_table	270884.53
5	furniture_decor	217031.99
6	cool_stuff	159989.99
7	auto	159486.70
8	housewares	150968.34
9	garden_tools	134342.27

	product_category_name_english	Q2 Sales Revenue
0	watches_gifts	411564.48
1	health_beauty	406278.63
2	bed_bath_table	330360.51
3	sports_leisure	274743.02
4	computers_accessories	272208.59
5	housewares	233452.95
6	furniture_decor	226599.59
7	auto	212387.53
8	cool_stuff	185065.37
9	garden_tools	163960.77

	product_category_name_english	Q3 Sales Revenue
0	health_beauty	374332.39
1	bed_bath_table	306673.96
2	watches_gifts	300140.10
3	sports_leisure	247473.30
4	housewares	200015.75
5	computers_accessories	194142.31
6	furniture_decor	189137.73
7	cool_stuff	173059.37
8	auto	146251.67
9	computers	118678.87

	product_category_name_english	Q4 Sales Revenue
0	watches_gifts	247899.86
1	bed_bath_table	199478.98
2	health_beauty	196706.55
3	sports_leisure	185092.79
4	toys	166138.64
5	computers_accessories	161076.08
6	cool_stuff	146522.40
7	furniture_decor	139726.85
8	garden_tools	104246.35
9	auto	100269.60

Proposal - How to improve FY19 Sales Revenue and Sales Amount

1. Maintain marketing resources on top line products:

`health_beauty`, `watches_gifts`, `bed_bath_table`, `sports_leisure`, `computers_accessories`, `furniture_decor`, `housewares`, `cool_stuff`, `auto`, `garden_tools`

1. Leverage marketing resources on potential product to archive 1 level up product hierarchy

`watches_gifts`, `computers_accessories`, `cool_stuff`

1. Leverage marketing resources seasonally on seasonal products:

Q3: `computers` ; Q4: `toys`;

Contribution

Name	Task
Liang XuChao	- Random forest regression analysis - Call apis - Delivery Time computation
Liao ZiXin	- Location, delivery time & product category analysis - Sales amount & revenue analysis - Proposal - FY19 Marketing plan
Lin Yan	- Data preparation - Linear Regression analysis - Proposal - FY19 Marketing plan

Reference

https://www.kaggle.com/need4data/olist-eda

https://www.kaggle.com/badamnarendra/olist-analysis-and-revenue-prediction

https://www.kaggle.com/rennatts/brazilian-e-commerce-analysis