Sas project

Olist电商平台数据分析

李瑞旻12110448 赵一菡12110248
刘　睿12110942 李思卉12110746

2024.05

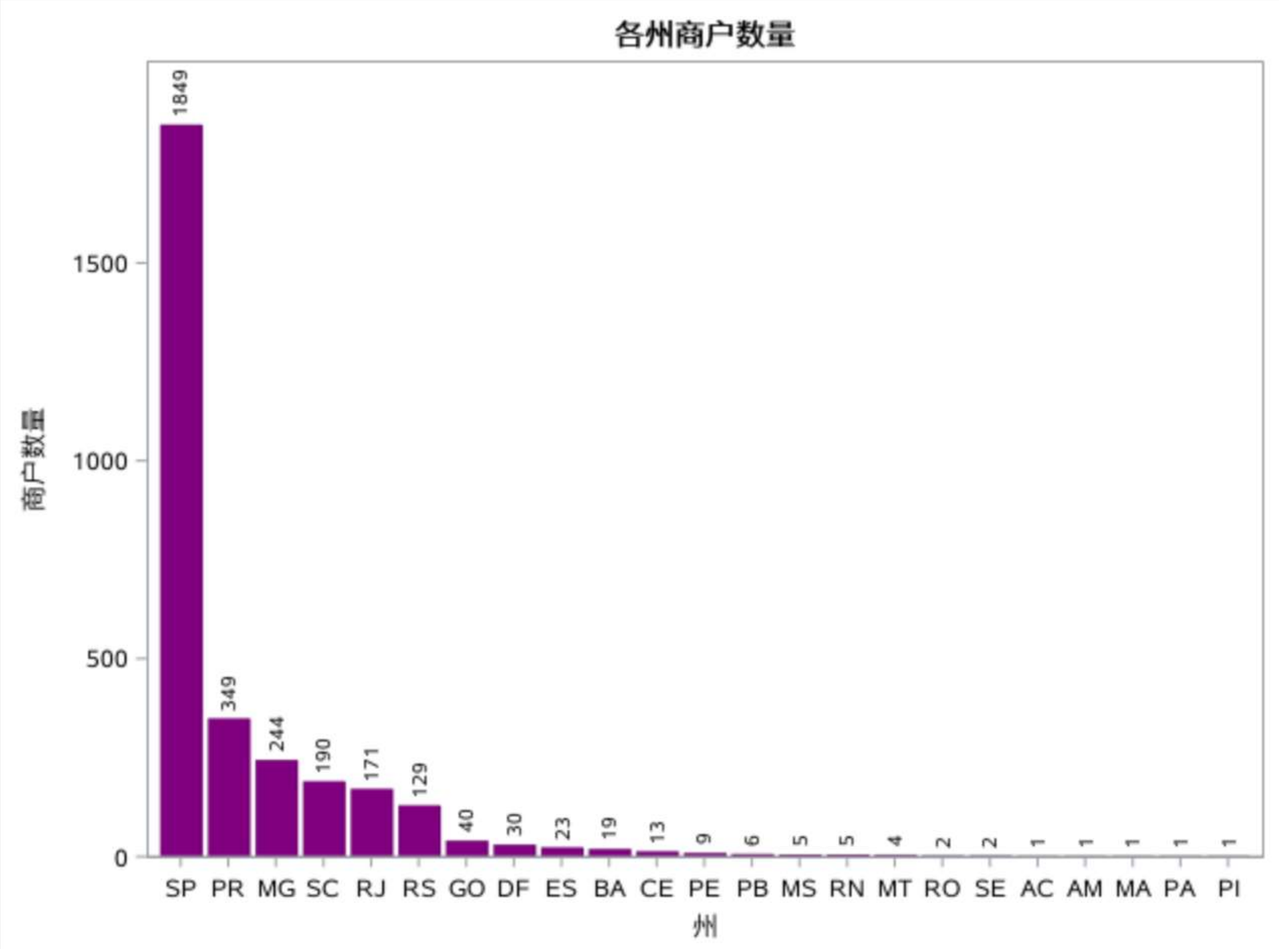
目录 Contents

01

商家和产品

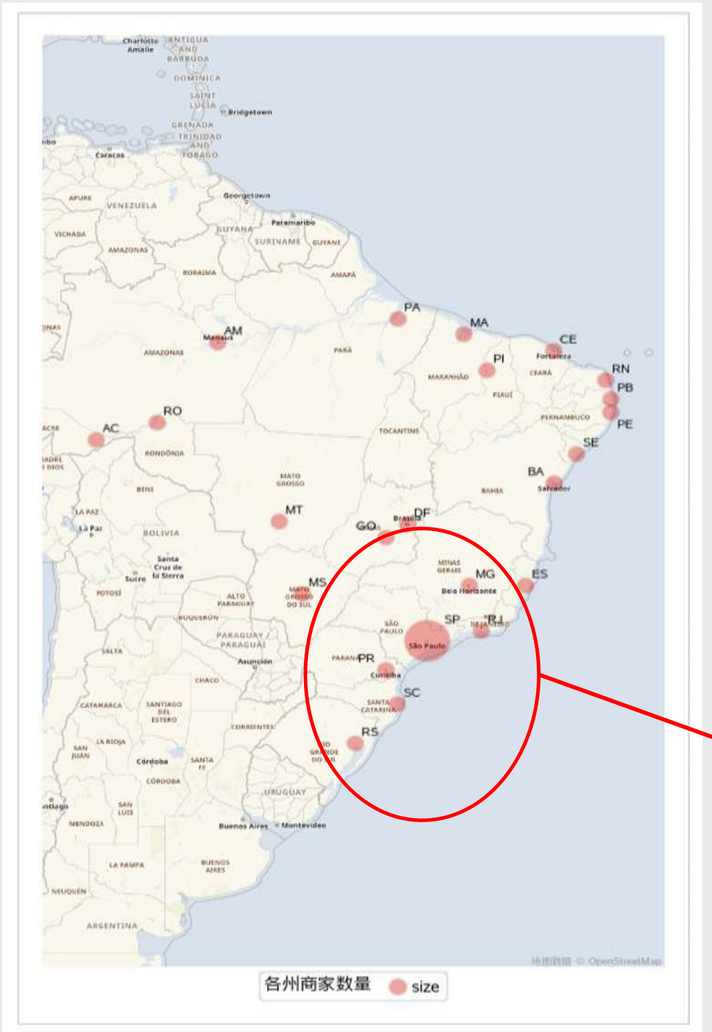
1.1 商家角度
1.2 产品角度

1.1.1 商家的地区分布



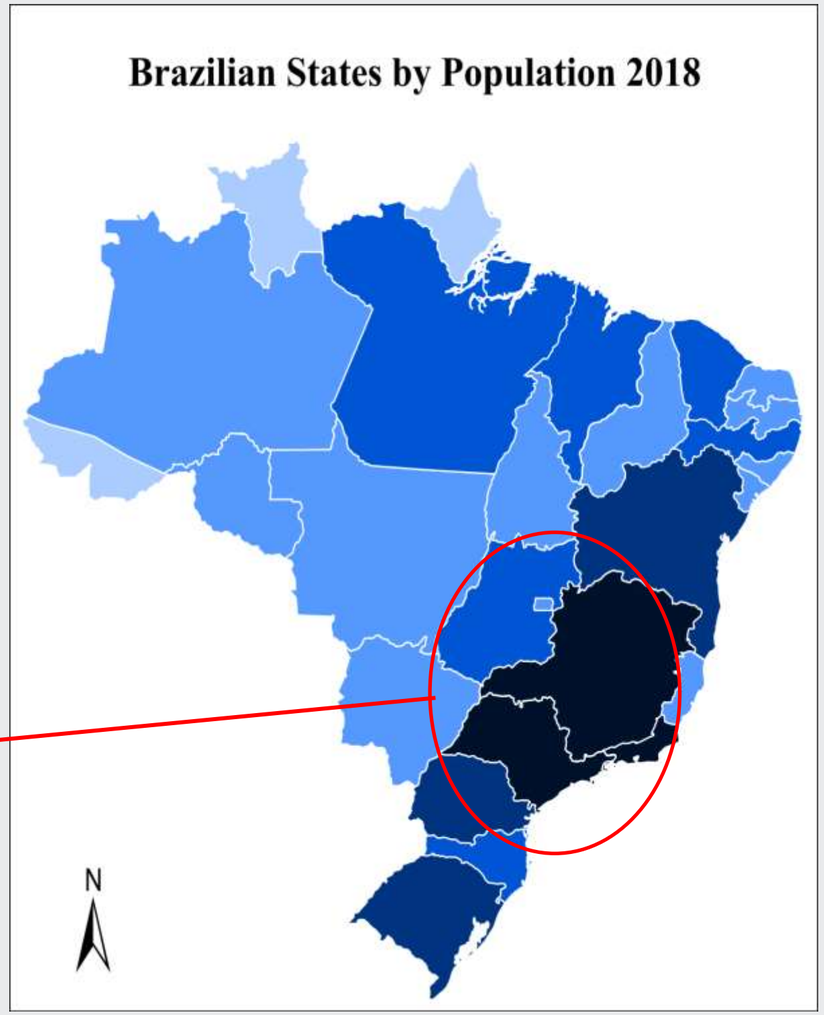
- SP-St.Paul
- PR- Paraná
- MG- Minas Gerais
- SC-Santa Catarina
- RJ-Rio de Janeiro

1.1.1 商家的地区分布

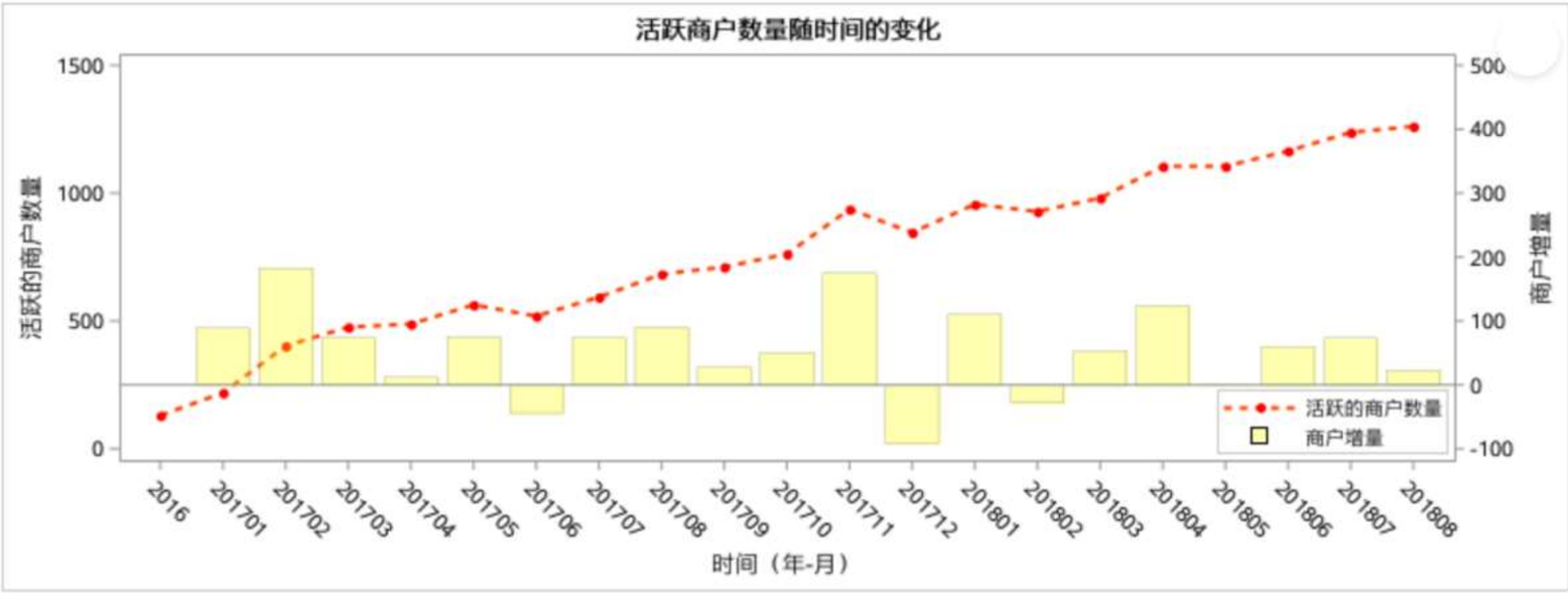


- Sellers distribution and population distribution

- southeast

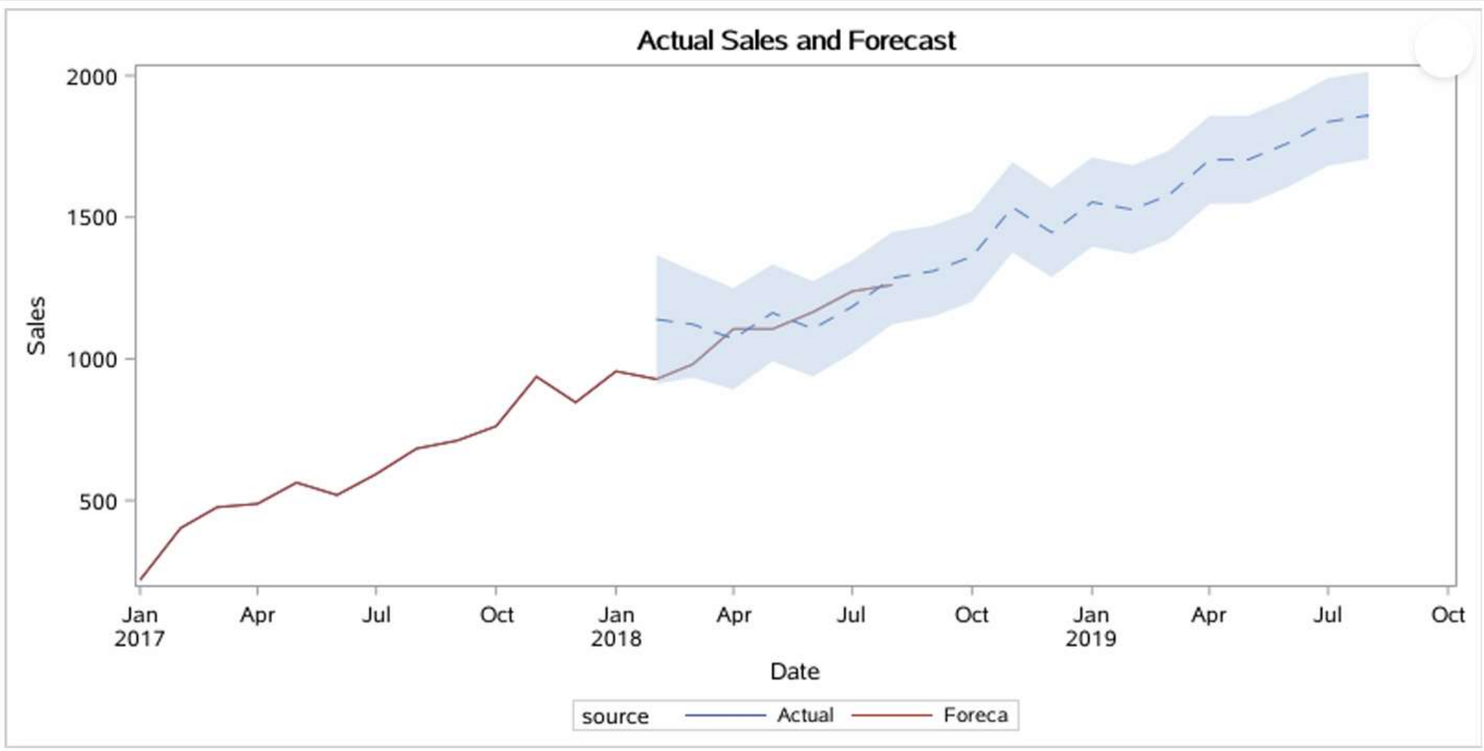


1.1.2 商家的时间变化



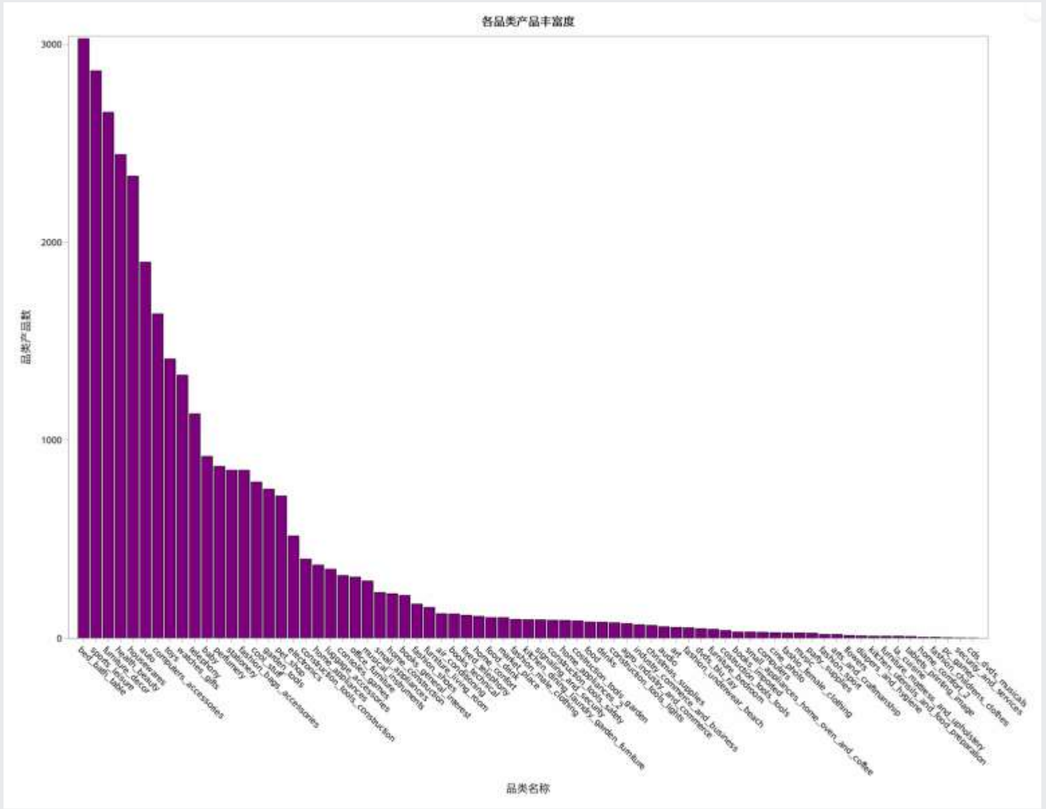
- The number of sellers active on the platform in each month is calculated based on the order time.
- Trend: increasing
- Increment: gradually flattened out

1.1.2 商家的时间变化

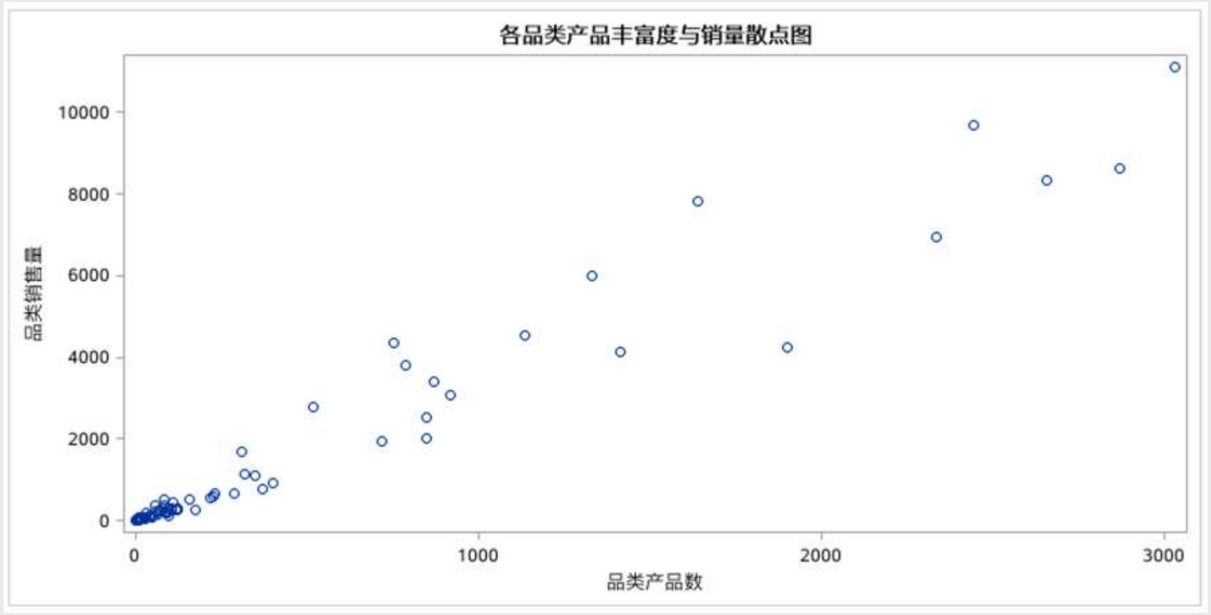


- proc ucm
- Time series forecast：Trend Seasonal Noise

1.2.1 品类产品丰富度

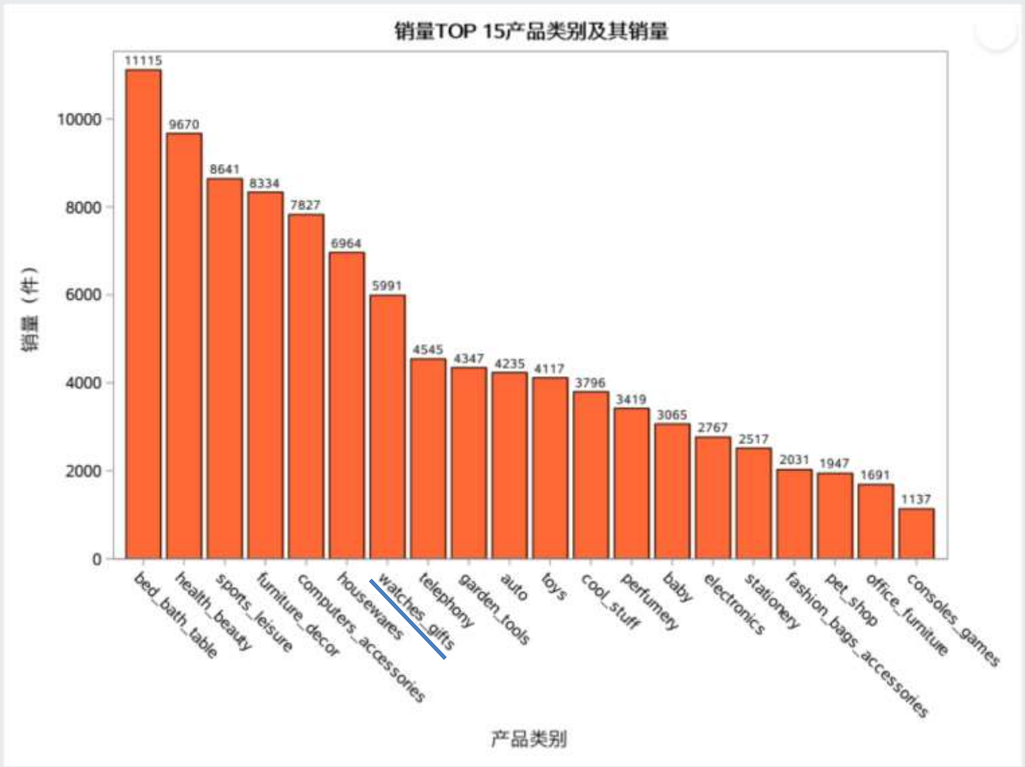


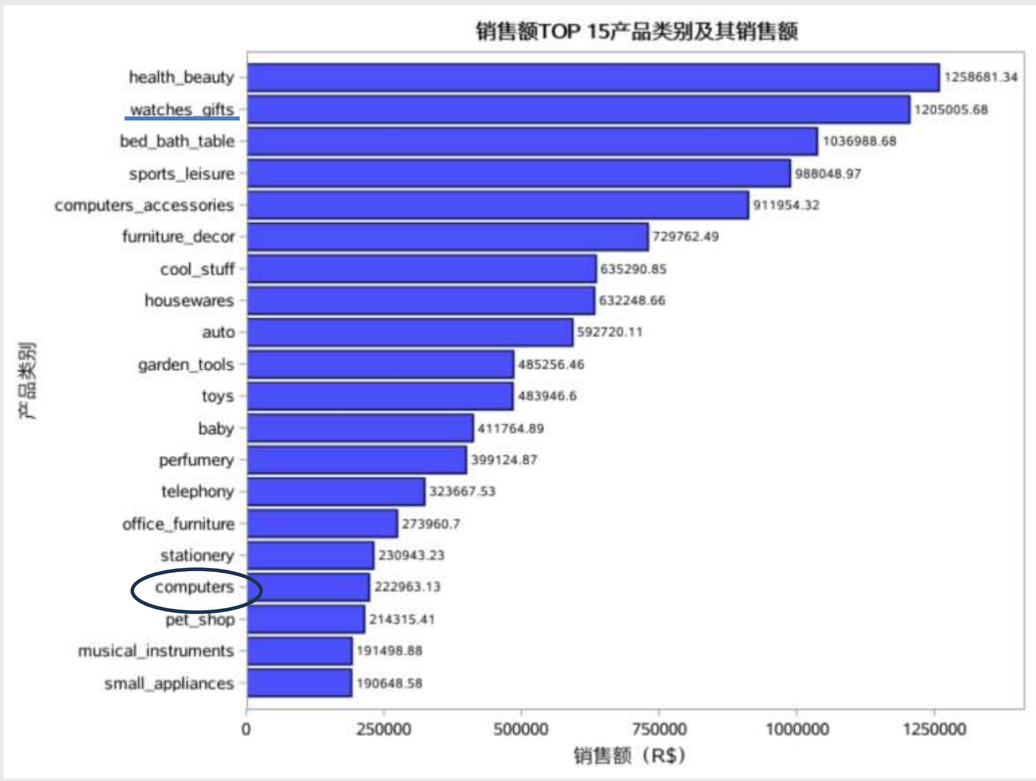
- 1~3029
- bed_bath_table，sports_leisure，furniture，health_beauty，housewares



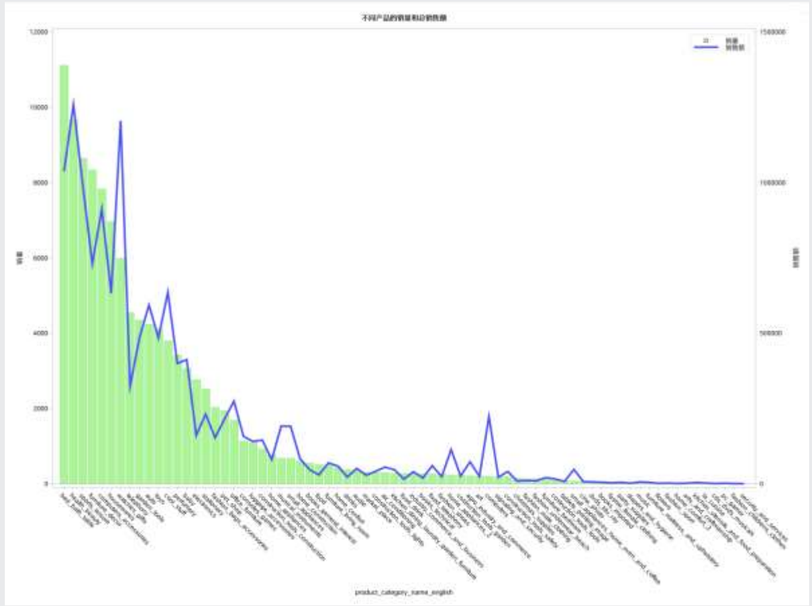
- Corr=0.97203
- Strongly positive correlation

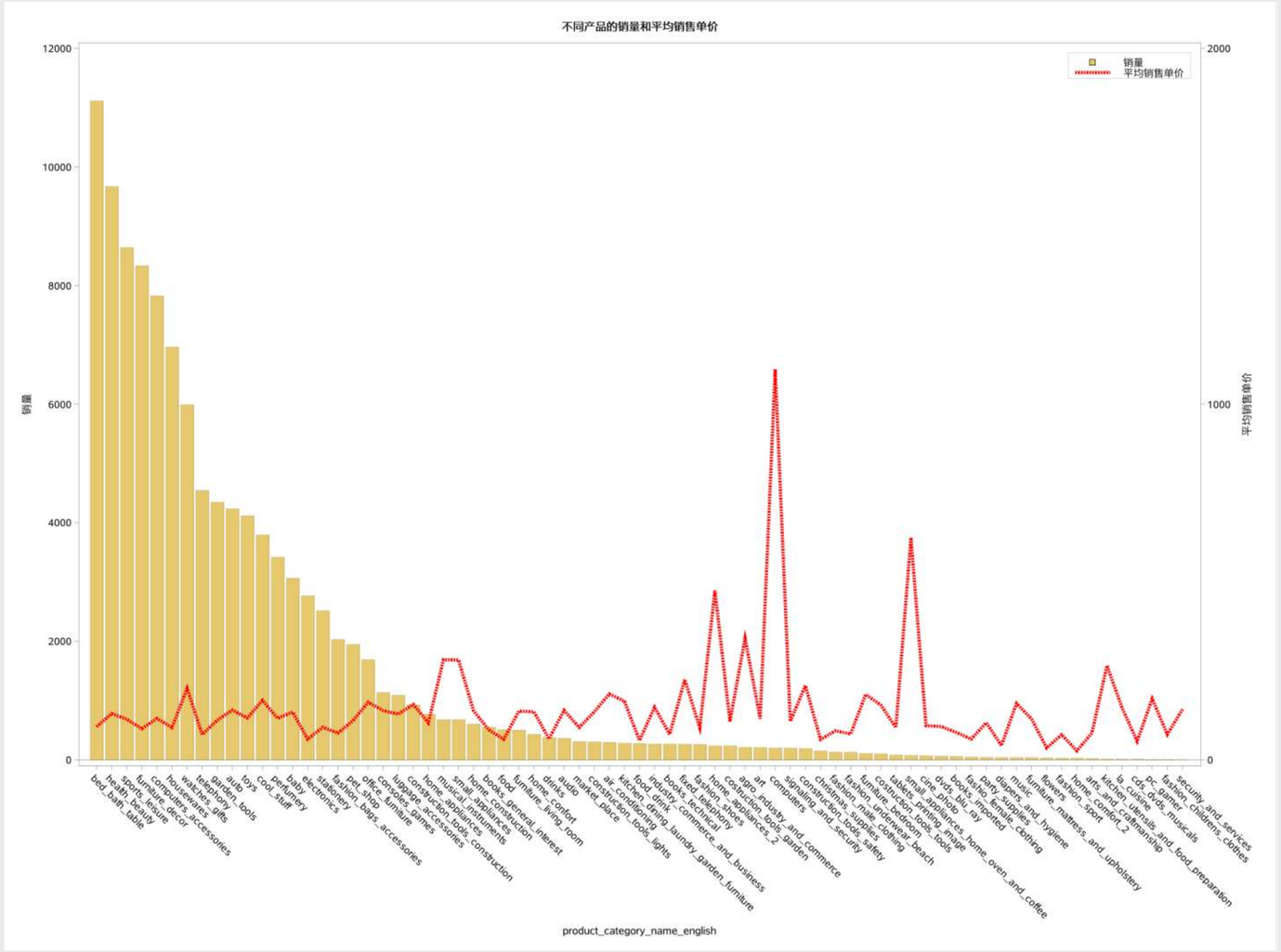
1.2.2 产品销量和销售额





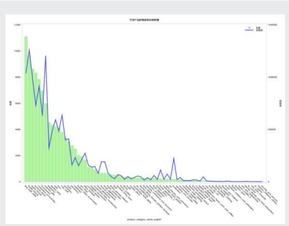
1.2.2 产品销量和销售额

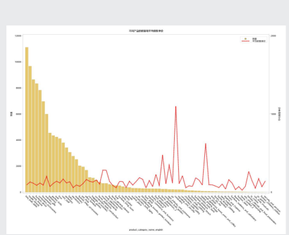




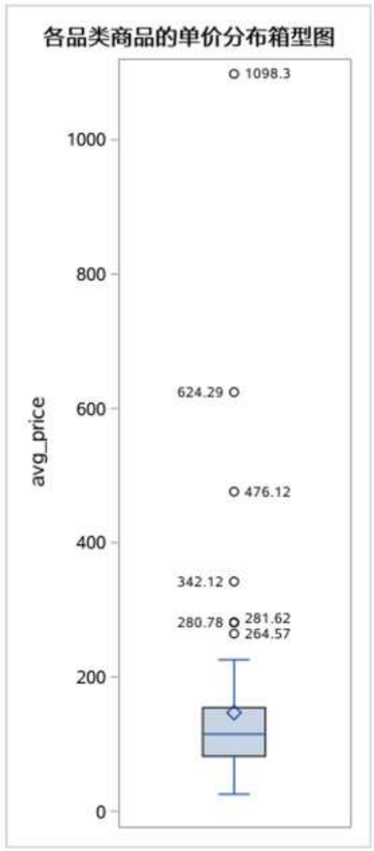
- Total：bed bath; health beauty
- Unit price：computer; small applications home oven and coffee

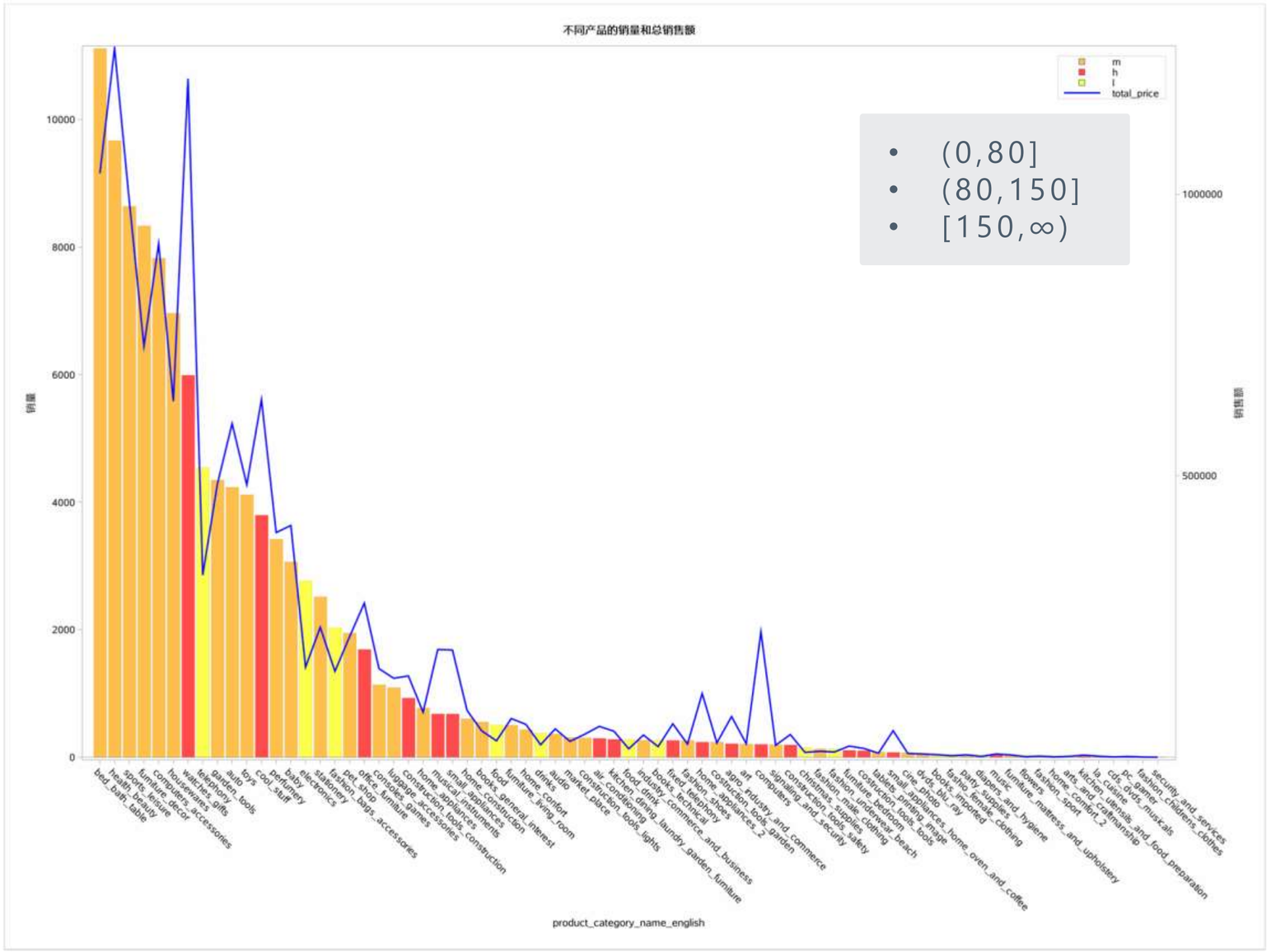
1.2.2 产品销量和销售额





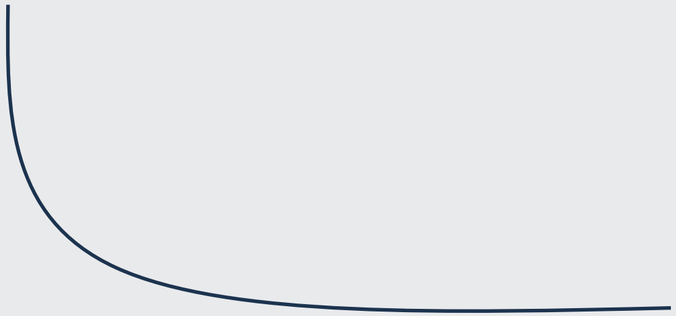
分位数（定义 5）	
水平	分位数
100% 最大值	1098.3405
99%	1098.3405
95%	342.1249
90%	225.6932
75% Q3	154.4073
50% 中位数	114.9495
25% Q1	81.8017
10%	57.9135
5%	52.1429
1%	25.3423
0% 最小值	25.3423





- (0,80]
- (80,150]
- [150,∞)

1.2.3 商品售出频次



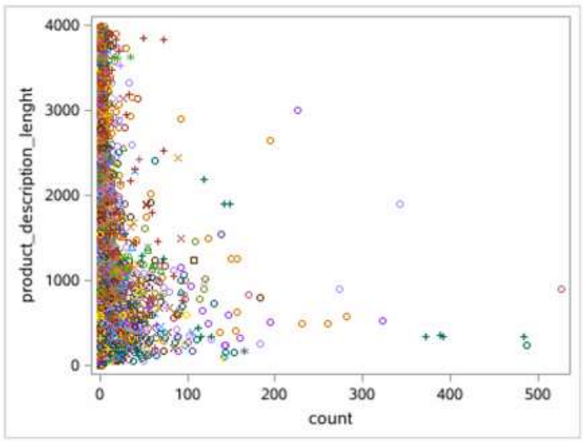
- Heavy-tailed
- Head Effect

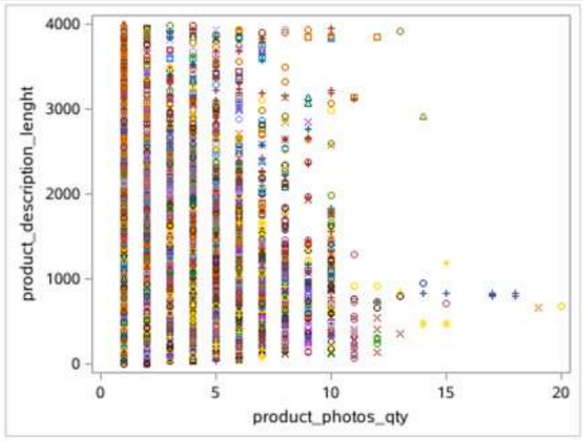
各售出次数的商品数

FREQ 过程

count	频数	百分比	累积频数	累积百分比
1	18117	54.98	18117	54.98
2	5817	17.65	23934	72.64
3	2651	8.05	26585	80.68
4	1534	4.66	28119	85.34
5	994	3.02	29113	88.35
6	736	2.23	29849	90.59
7	515	1.56	30364	92.15
8	379	1.15	30743	93.30
9	289	0.88	31032	94.18
10	251	0.76	31283	94.94
11	185	0.56	31468	95.50
12	156	0.47	31624	95.97
13	130	0.39	31754	96.37
194	1	0.00	32936	99.95
195	1	0.00	32937	99.96
197	1	0.00	32938	99.96
226	1	0.00	32939	99.96
231	1	0.00	32940	99.97
260	1	0.00	32941	99.97
274	1	0.00	32942	99.97
281	1	0.00	32943	99.98
323	1	0.00	32944	99.98
343	1	0.00	32945	99.98
373	1	0.00	32946	99.98
388	1	0.00	32947	99.99
392	1	0.00	32948	99.99
484	1	0.00	32949	99.99
488	1	0.00	32950	100.00
527	1	0.00	32951	100.00

1.2.4 销量预测和建议





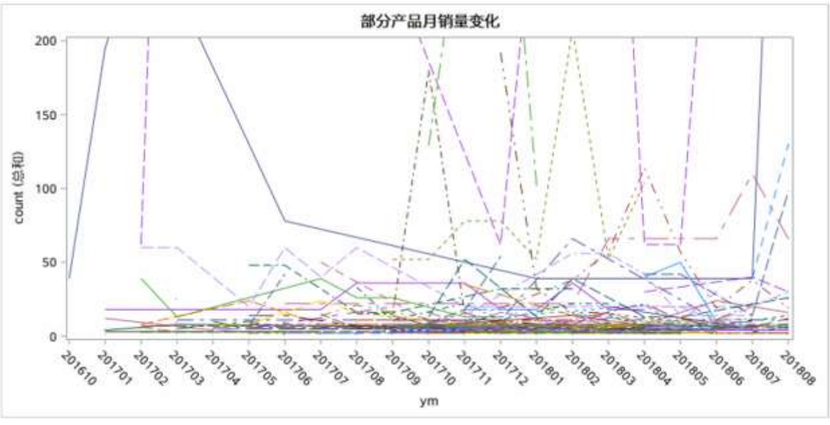
销量和几个变量的相关系数

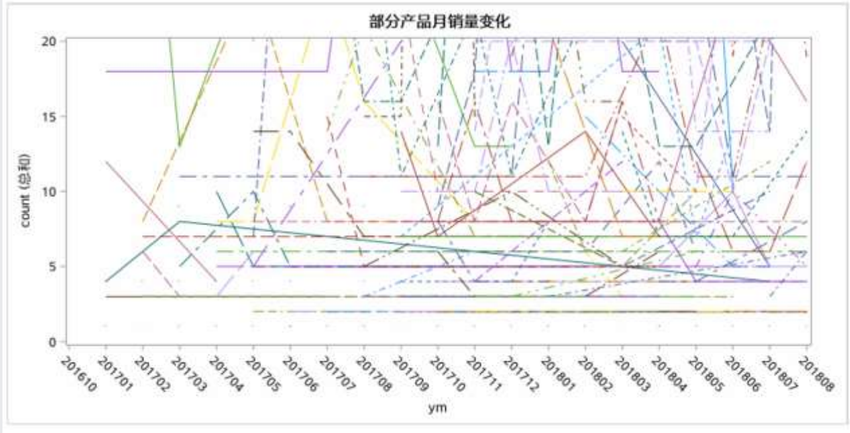
CORR 过程

Pearson 相关系数, N = 32341				
	count	avg_price	product_photos_qty	product_description_lenght
count	1.00000	-0.03281	0.00385	0.00831
avg_price	-0.03281	1.00000	0.06008	0.19265
product_photos_qty	0.00385	0.06008	1.00000	0.10874
product_description_lenght	0.00831	0.19265	0.10874	1.00000

- Monthly？
- Transform？

1.2.4 销量预测和建议





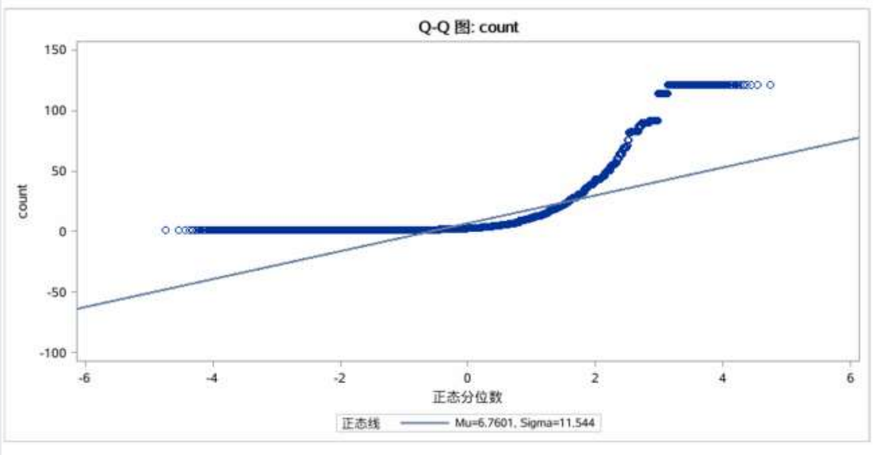
- Leveled
- Last month

CORR 过程

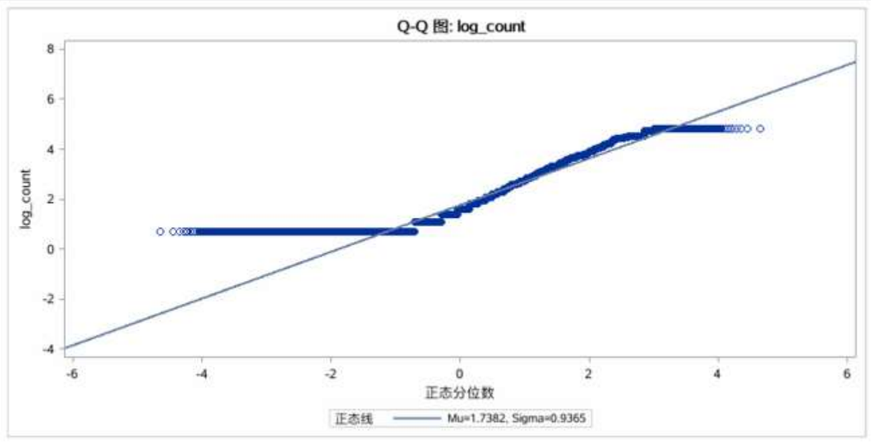
Pearson 相关系数, N = 581904

	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm	price	review_score	freight_value	count	avg_price	avg_freight	last_ym_count
product_name_lenght	1.00000	0.06873	0.10961	-0.00173	0.05261	-0.03515	0.09904	-0.00368	-0.00875	0.00967	0.05018	-0.00884	0.00605	0.05038
product_description_lenght	0.06873	1.00000	0.10724	0.09520	-0.02367	0.07639	-0.11896	0.20482	0.01796	0.13025	-0.02829	0.20334	0.13589	-0.02847
product_photos_qty	0.10961	0.10724	1.00000	-0.01037	0.01557	-0.09423	-0.03029	0.03066	0.02488	-0.03006	0.02560	0.03188	-0.02416	0.02598
product_weight_g	-0.00173	0.09520	-0.01037	1.00000	0.42798	0.60497	0.47151	0.33573	-0.02172	0.58734	-0.02971	0.33617	0.64543	-0.02906
product_length_cm	0.05261	-0.02367	0.01557	0.42798	1.00000	0.18111	0.56899	0.13438	-0.01797	0.24810	0.01949	0.13064	0.28124	0.02009
product_height_cm	-0.03515	0.07639	-0.09423	0.60497	0.18111	1.00000	0.24064	0.20960	-0.02666	0.37182	-0.01557	0.20878	0.41357	-0.01541
product_width_cm	0.09904	-0.11896	-0.03029	0.47151	0.56899	0.24064	1.00000	0.15817	-0.00662	0.27038	0.05906	0.15403	0.30307	0.05942
price	-0.00368	0.20482	0.03066	0.33573	0.13438	0.20960	0.15817	1.00000	0.00814	0.38266	-0.05474	0.99053	0.41395	-0.05344
review_score	-0.00875	0.01796	0.02488	-0.02172	-0.01797	-0.02666	-0.00662	0.00814	1.00000	-0.03384	-0.01520	0.00699	-0.00813	-0.01507
freight_value	0.00967	0.13025	-0.03006	0.58734	0.24810	0.37182	0.27038	0.38266	-0.03384	1.00000	-0.04024	0.38049	0.66470	-0.03954
count	0.05018	-0.02829	0.02560	-0.02971	0.01949	-0.01557	0.05906	-0.05474	-0.01520	-0.04024	1.00000	-0.05992	-0.05761	0.97760
avg_price	-0.00884	0.20334	0.03188	0.33617	0.13064	0.20878	0.15403	0.99053	0.00699	0.38049	-0.05992	1.00000	0.41381	-0.05804
avg_freight	0.00605	0.13589	-0.02416	0.64543	0.28124	0.41357	0.30307	0.41395	-0.00813	0.66470	-0.05761	0.41381	1.00000	-0.05635
last_ym_count	0.05038	-0.02847	0.02598	-0.02906	0.02009	-0.01541	0.05942	-0.05344	-0.01507	-0.03954	0.97760	-0.05804	-0.05635	1.00000

1.2.4 销量预测和建议



Log-transform



CORR 过程

Pearson 相关系数, N = 581904

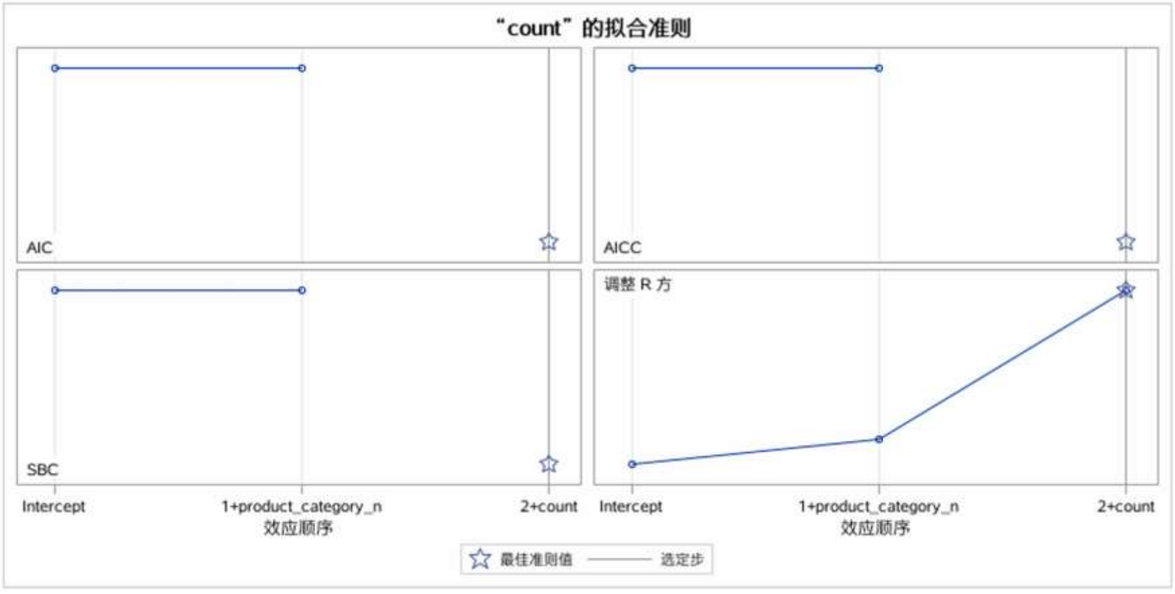
	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm	price	review_score	freight_value	count	avg_price	avg_freight	last_ym_count
product_name_lenght	1.00000	0.06873	0.10961	-0.00173	0.05261	-0.03515	0.09904	-0.00368	-0.00875	0.00967	0.05785	-0.00884	0.00605	0.05838
product_description_lenght	0.06873	1.00000	0.10724	0.09520	-0.02367	0.07639	-0.11896	0.20482	0.01796	0.13025	-0.01463	0.20334	0.13589	-0.01545
product_photos_qty	0.10961	0.10724	1.00000	-0.01037	0.01557	-0.09423	-0.03029	0.03066	0.02488	-0.03006	-0.03754	0.03188	-0.02416	-0.03639
product_weight_g	-0.00173	0.09520	-0.01037	1.00000	0.42798	0.60497	0.47151	0.33573	-0.02172	0.58734	-0.04359	0.33617	0.64543	-0.04177
product_length_cm	0.05261	-0.02367	0.01557	0.42798	1.00000	0.18111	0.56899	0.13438	-0.01797	0.24810	-0.02409	0.13064	0.28124	-0.02249
product_height_cm	-0.03515	0.07639	-0.09423	0.60497	0.18111	1.00000	0.24064	0.20960	-0.02666	0.37182	-0.01739	0.20878	0.41357	-0.01716
product_width_cm	0.09904	-0.11896	-0.03029	0.47151	0.56899	0.24064	1.00000	0.15817	-0.00662	0.27038	0.03474	0.15403	0.30307	0.03553
price	-0.00368	0.20482	0.03066	0.33573	0.13438	0.20960	0.15817	1.00000	0.00814	0.38266	-0.07081	0.99053	0.41395	-0.06601
review_score	-0.00875	0.01796	0.02488	-0.02172	-0.01797	-0.02666	-0.00662	0.00814	1.00000	-0.03384	-0.01855	0.00699	-0.00813	-0.01787
freight_value	0.00967	0.13025	-0.03006	0.58734	0.24810	0.37182	0.27038	0.38266	-0.03384	1.00000	-0.04163	0.38049	0.66470	-0.03916
count	0.05785	-0.01463	-0.03754	-0.04359	-0.02409	-0.01739	0.03474	-0.07081	-0.01855	-0.04163	1.00000	-0.07632	-0.04588	0.91394
avg_price	-0.00884	0.20334	0.03188	0.33617	0.13064	0.20878	0.15403	0.99053	0.00699	0.38049	-0.07632	1.00000	0.41381	-0.07039
avg_freight	0.00605	0.13589	-0.02416	0.64543	0.28124	0.41357	0.30307	0.41395	-0.00813	0.66470	-0.04588	0.41381	1.00000	-0.04311
last_ym_count	0.05838	-0.01545	-0.03639	-0.04177	-0.02249	-0.01716	0.03553	-0.06601	-0.01787	-0.03916	0.91394	-0.07039	-0.04311	1.00000

1.2.4 销量预测和建议

GLMSELECT 过程
选定模型

选定模型是最后一步（第 2 步）处的精确拟合模型。

效应:	Intercept count product_category_nam



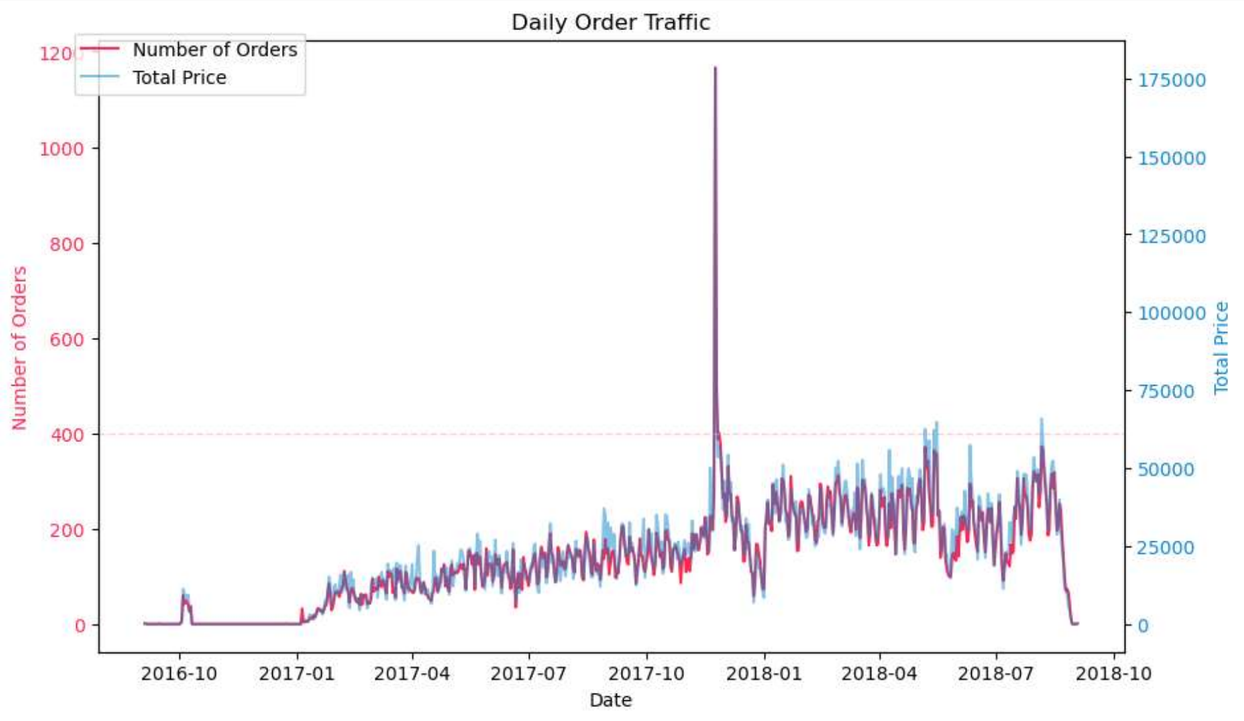
Suggestion:

- Offer more variable products
- Focus on top category products
- Improve product descriptions and scoring function

02

客户角度

订单量&订单金额

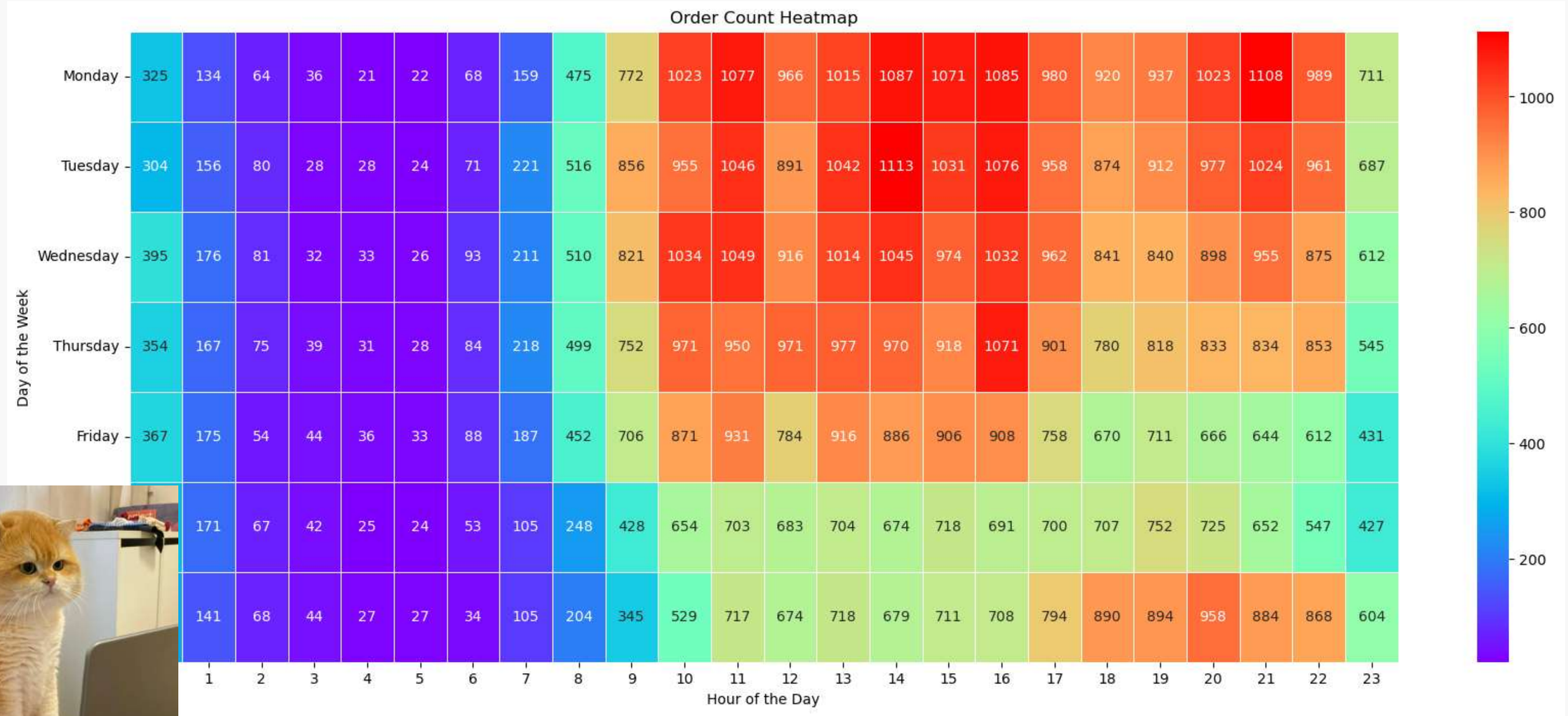


黑色星期五！


```
daily_orders[daily_orders > 500].index
```

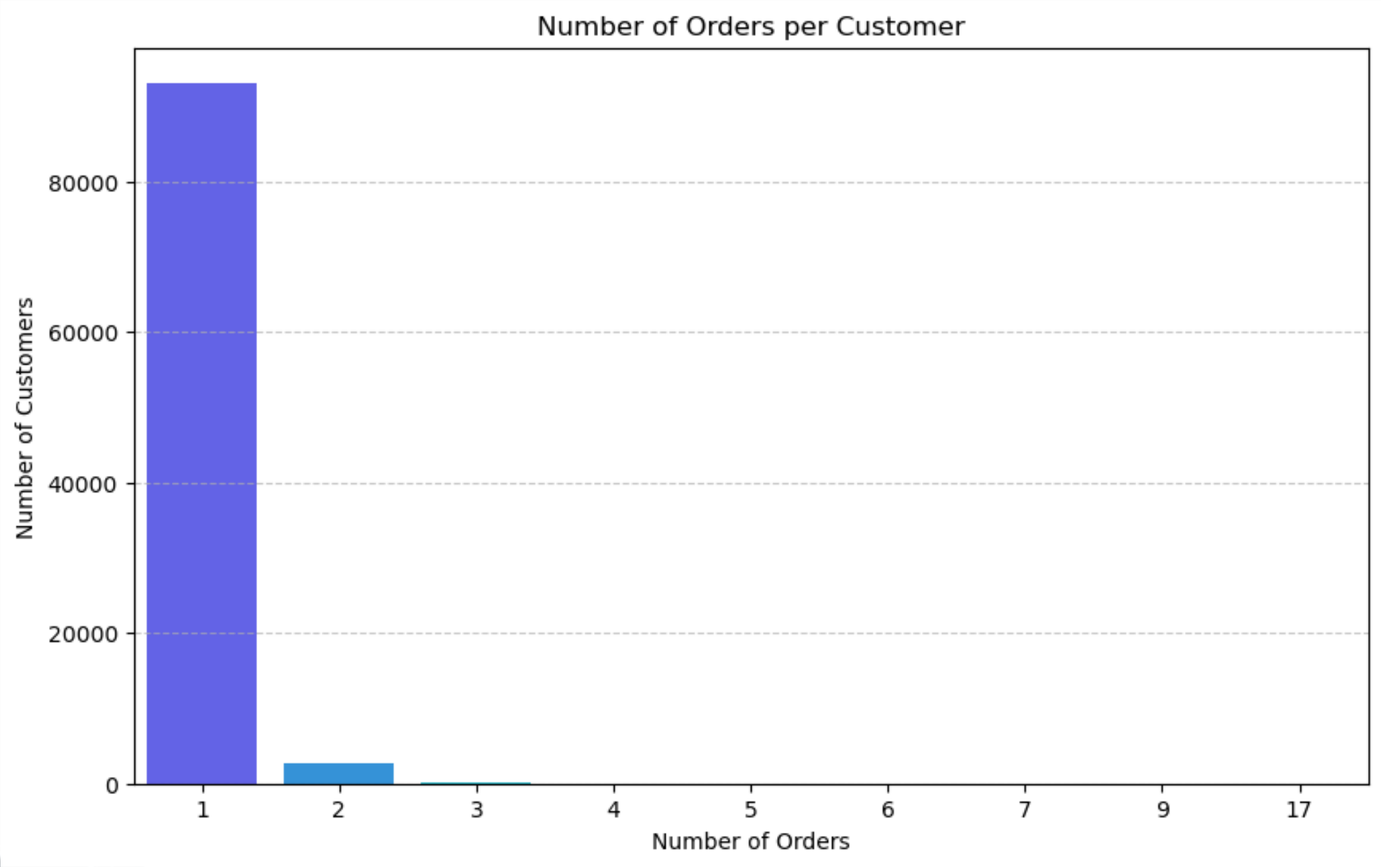
```
DatetimeIndex(['2017-11-24'], dtype='datetime64[ns]', name='order_purchase_timestamp', freq='D')
```

订单量&订单金额


Order Count Heatmap
Monday 325 134 64 36 21 22 68 159 475 772 1023 1077 966 1015 1087 1071 1085 980 920 937 1023 1108 989 711
Tuesday 304 156 80 28 28 24 71 221 516 856 955 1046 891 1042 1113 1031 1076 958 874 912 977 1024 961 687
Wednesday 395 176 81 32 33 26 93 211 510 821 1034 1049 916 1014 1045 974 1032 962 841 840 898 955 875 612
Thursday 354 167 75 39 31 28 84 218 499 752 971 950 971 977 970 918 1071 901 780 818 833 834 853 545
Friday 367 175 54 44 36 33 88 187 452 706 871 931 784 916 886 906 908 758 670 711 666 644 612 431
171 67 42 25 24 53 105 248 428 654 703 683 704 674 718 691 700 707 752 725 652 547 427
141 68 44 27 27 34 105 204 345 529 717 674 718 679 711 708 794 890 894 958 884 868 604
Day of the Week
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of the Day
1000
800
600
400
200

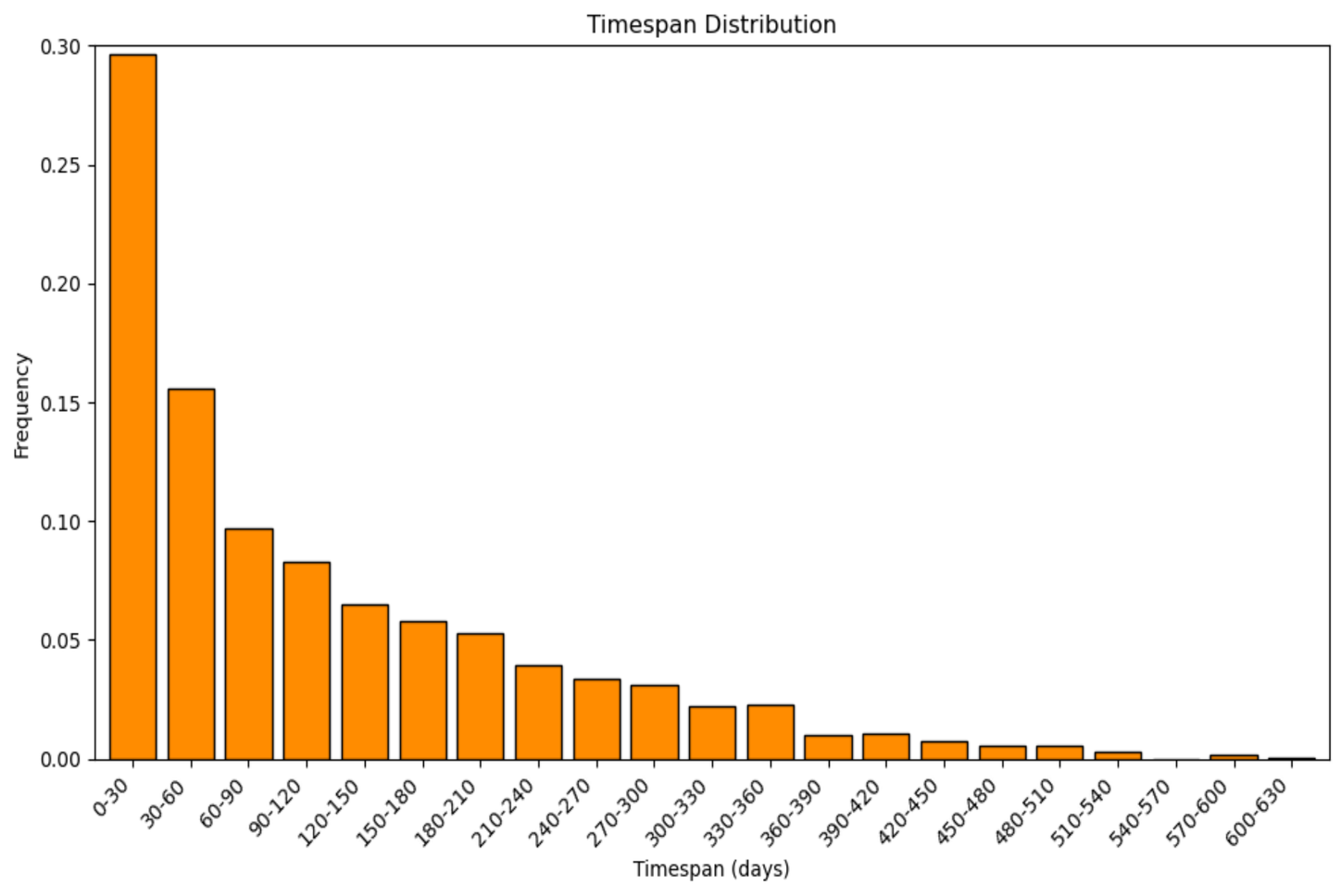

摸鱼时间到!

复购情况



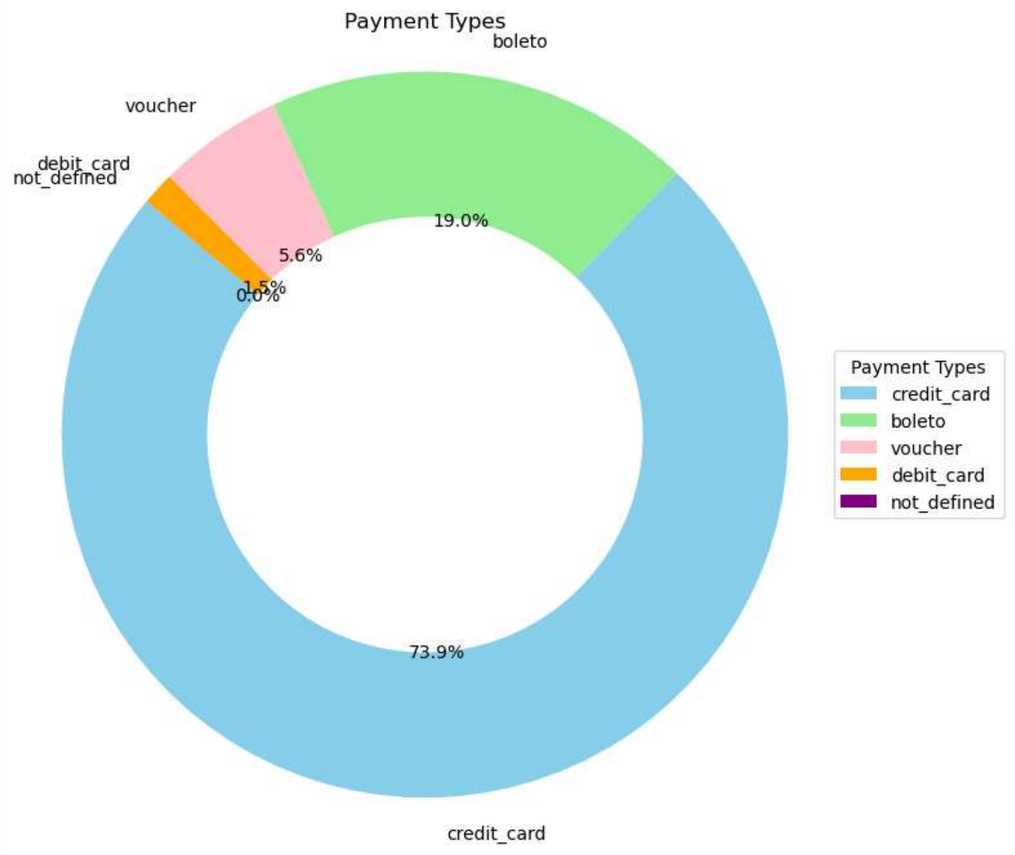
- 购买1次后继续进行购买的比例：3.12%
- 购买2次后继续进行购买的比例：8.41 %
- 购买3次后继续进行购买的比例：19.44 %

复购行为时间差



第1次和第2次购买时间差

支付途径

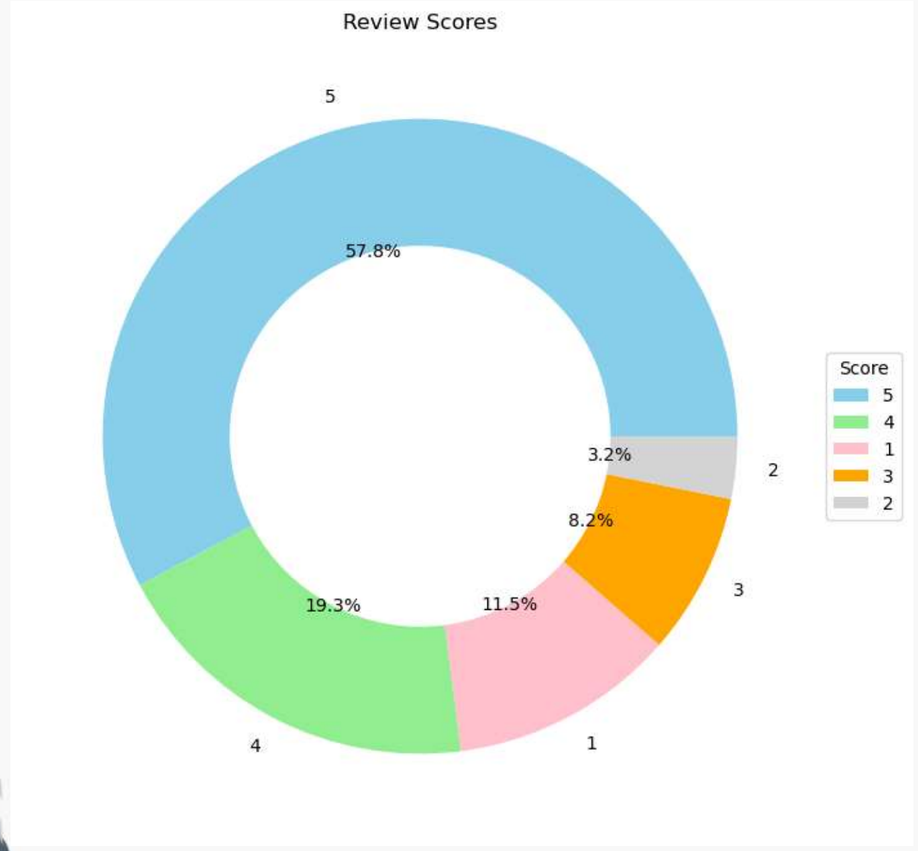

Payment Types
boleto
voucher
debit_card
not_defined
19.0%
5.6%
1.5%
0.0%
73.9%
credit_card
Payment Types
credit_card
boleto
voucher
debit_card
not_defined

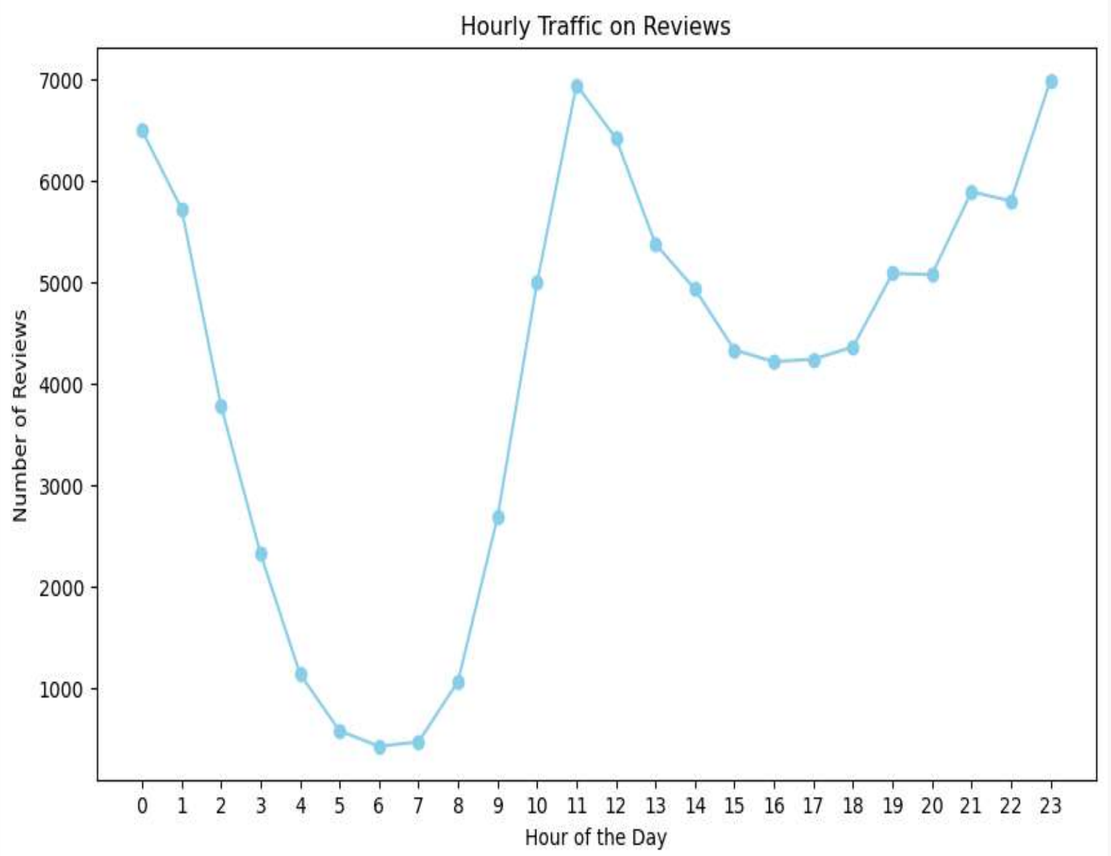
分期方式影响因素

```
> ModelInteract <- lm(payment_installments ~ payment_value + payment_type + payment_value * payment_type, data = payments)
> summary(ModelInteract)

Call:
lm(formula = payment_installments ~ payment_value + payment_type +
    payment_value * payment_type, data = payments)

Residuals:
    Min      1Q  Median      3Q     Max
-67.750  -1.816   0.000   0.246  20.115

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          1.000e+00  1.952e-02   51.24   <2e-16 ***
payment_value                       -4.808e-16  7.560e-05    0.00        1
payment_typecredit_card              1.718e+00  2.201e-02   78.06   <2e-16 ***
payment_typedebit_card              -1.558e-13  6.992e-02    0.00        1
payment_typevoucher                 -1.698e-13  3.953e-02    0.00        1
payment_value:payment_typecredit_card  4.833e-03  8.412e-05   57.45   <2e-16 ***
payment_value:payment_typedebit_card   3.629e-16  2.482e-04    0.00        1
payment_value:payment_typevoucher      3.067e-16  2.695e-04    0.00        1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.271 on 103875 degrees of freedom
Multiple R-squared:  0.2858,    Adjusted R-squared:  0.2857
F-statistic:  5938 on 7 and 103875 DF,  p-value: < 2.2e-16
```

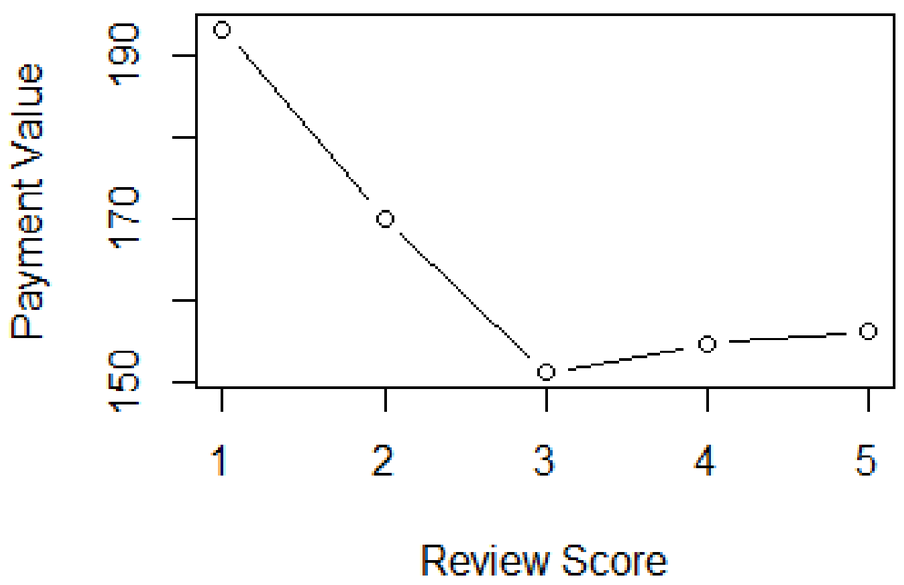
```
> anova(ModelPriceType, ModelInteract)
Analysis of Variance Table

Model 1: payment_installments ~ payment_value + payment_type
Model 2: payment_installments ~ payment_value + payment_type + payment_value *
    payment_type
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1 103878 555223
2 103875 535713  3     19511 1261 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

用户评分


Review Scores
5
57.8%
3.2%
2
8.2%
3
19.3%
11.5%
4
1
Score
5
4
1
3
2

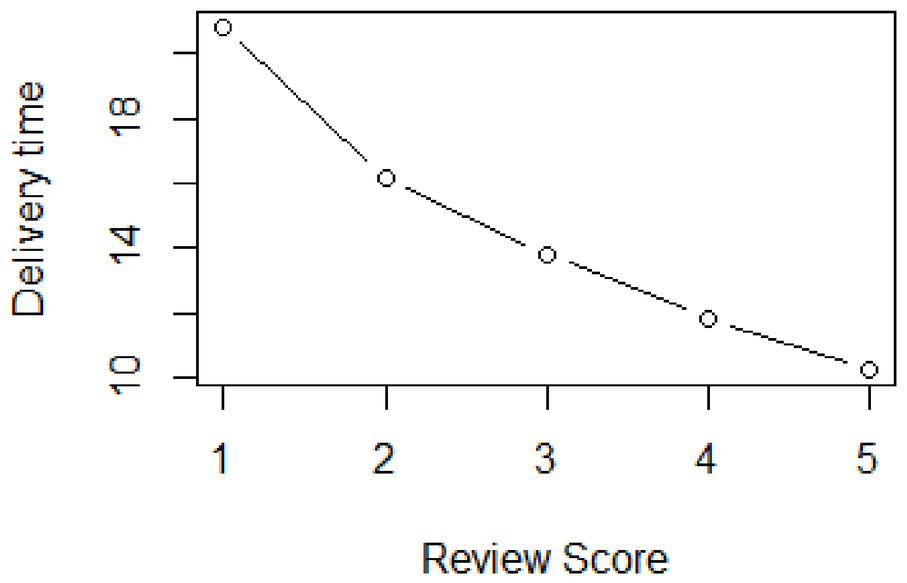

Hourly Traffic on Reviews
Number of Reviews
7000
6000
5000
4000
3000
2000
1000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of the Day

用户评分影响因素——订单金额




```
> summary(ModelValue)

Call:
lm(formula = review_score ~ payment_value, data = score)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1920 -0.1820  0.8164  0.8345  2.5314

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.195e+00  5.131e-03  817.59   <2e-16 ***
payment_value -2.492e-04  1.905e-05  -13.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.284 on 96356 degrees of freedom
Multiple R-squared:  0.001773,  Adjusted R-squared:  0.001763
F-statistic: 171.2 on 1 and 96356 DF,  p-value: < 2.2e-16
```

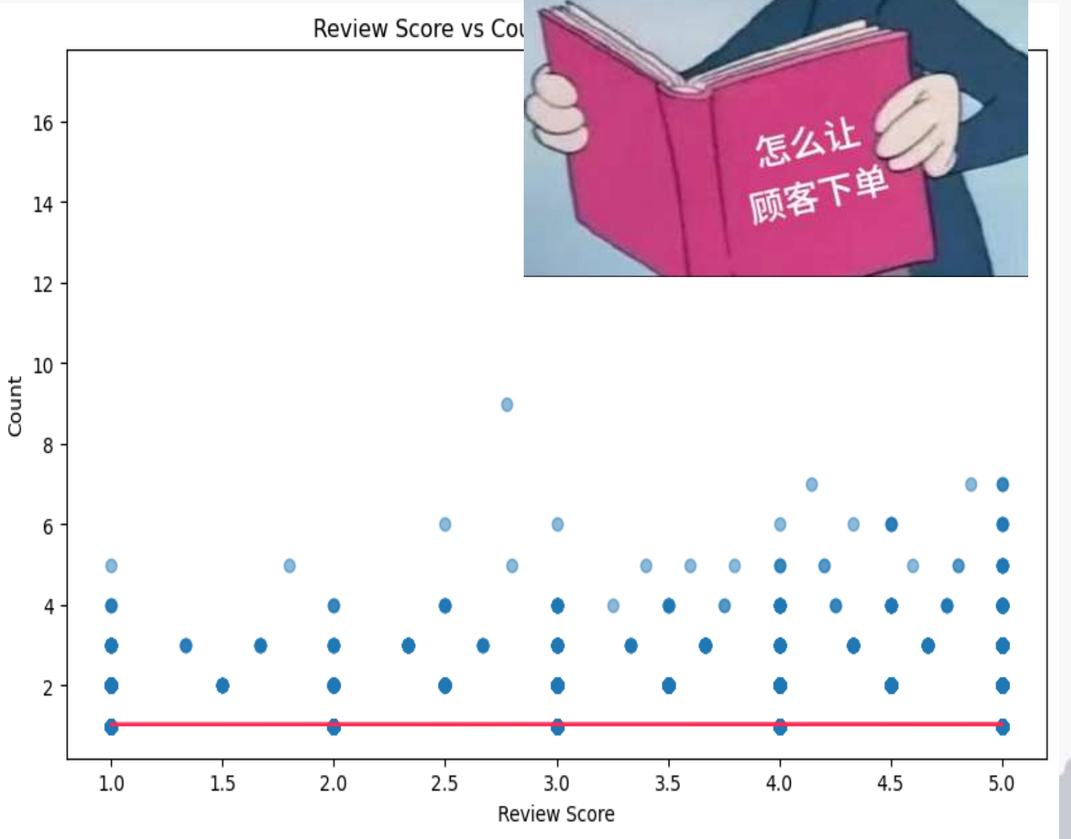
用户评分影响因素——收货时间




```
> cor(score$review_score, score$deliver_time)
[1] -0.3336026
> ModelTime <- lm(review_score ~ deliver_time, data = score)
> summary(ModelTime)

Call:
lm(formula = review_score ~ deliver_time, data = score)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7019 -0.4301  0.5246  0.7511  8.7705

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.7019397  0.0063221   743.7   <2e-16 ***
deliver_time -0.0453069  0.0004125  -109.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.211 on 96356 degrees of freedom
Multiple R-squared:  0.1113,    Adjusted R-squared:  0.1113
F-statistic: 1.207e+04 on 1 and 96356 DF,  p-value: < 2.2e-16
```

用户评分对复购行为的影响


```
> Model <- lm(score_reorder$count ~ score_reorder$review_score)
> summary(Model)

Call:
lm(formula = score_reorder$count ~ score_reorder$review_score)

Residuals:
    Min      1Q  Median      3Q     Max
-0.0411 -0.0411 -0.0411 -0.0394 15.9591

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.0369227  0.0026456 391.939   <2e-16 ***
score_reorder$review_score 0.0008272  0.0006153   1.344    0.179
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2549 on 95378 degrees of freedom
Multiple R-squared:  1.895e-05,  Adjusted R-squared:  8.466e-06
F-statistic: 1.807 on 1 and 95378 DF,  p-value: 0.1788
```



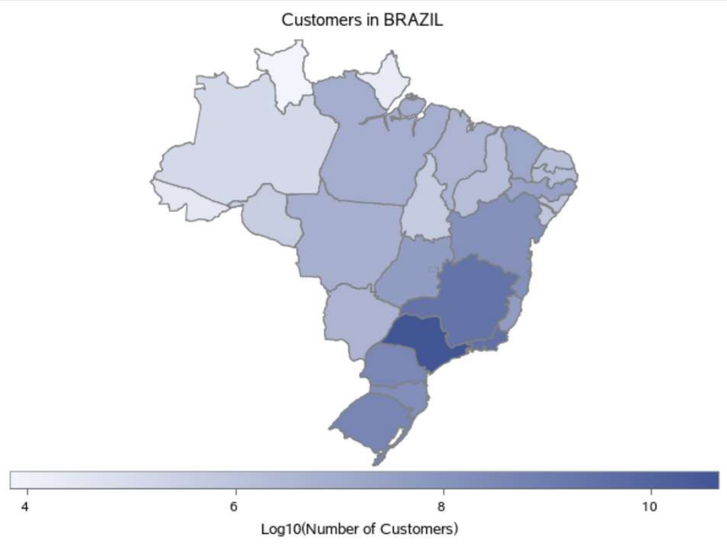
03

物流情况

2 main idea

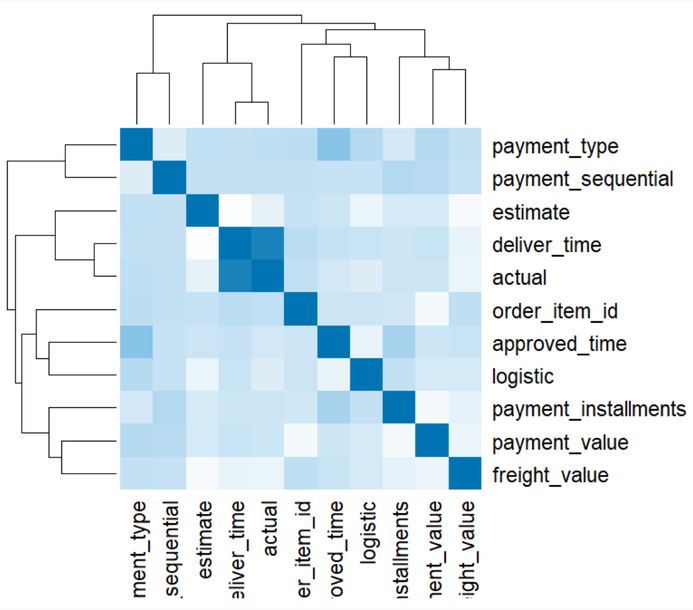
1st

Geolocation distribution

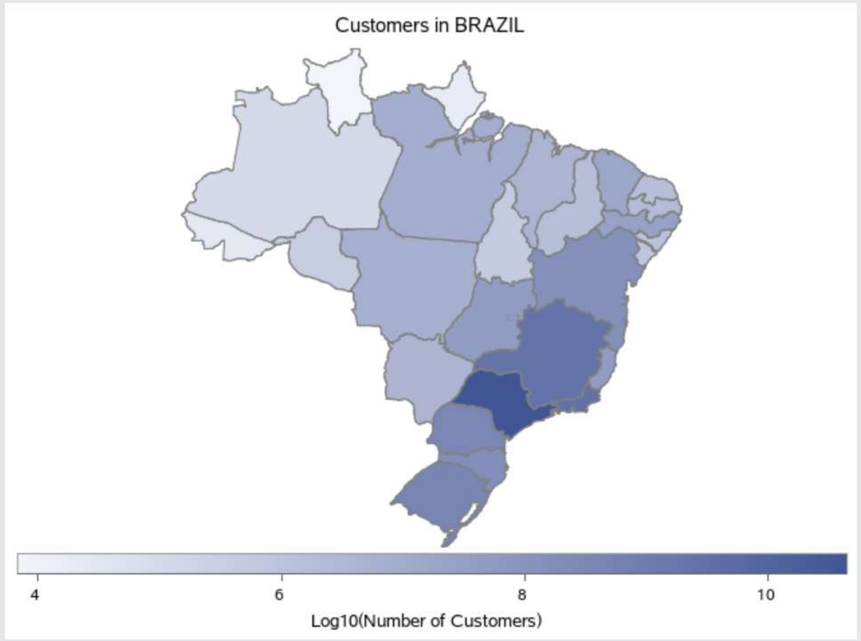


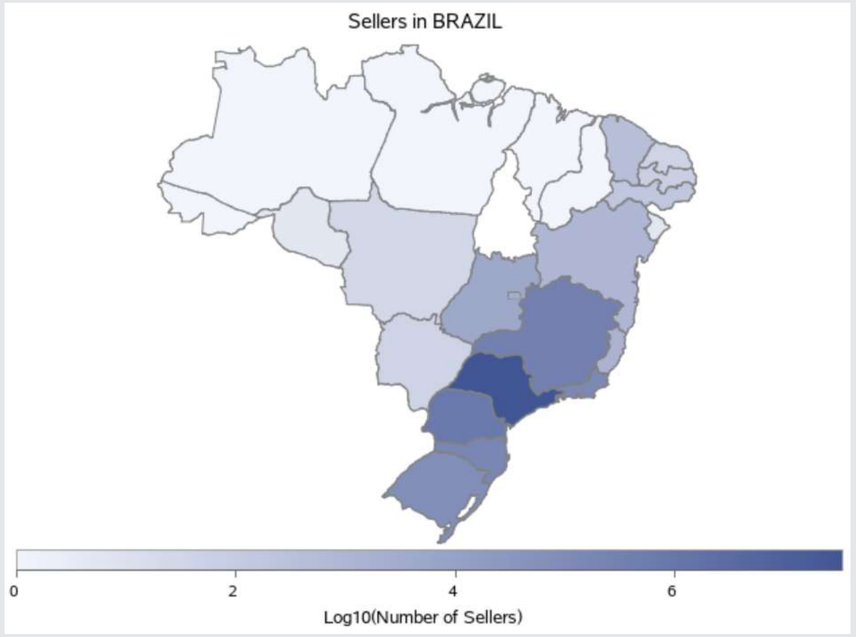
2nd

order time

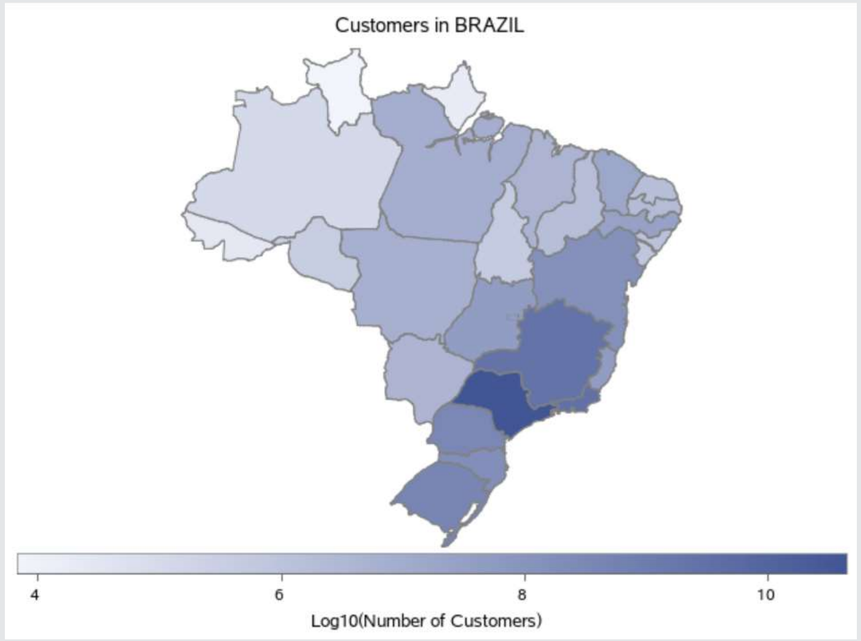


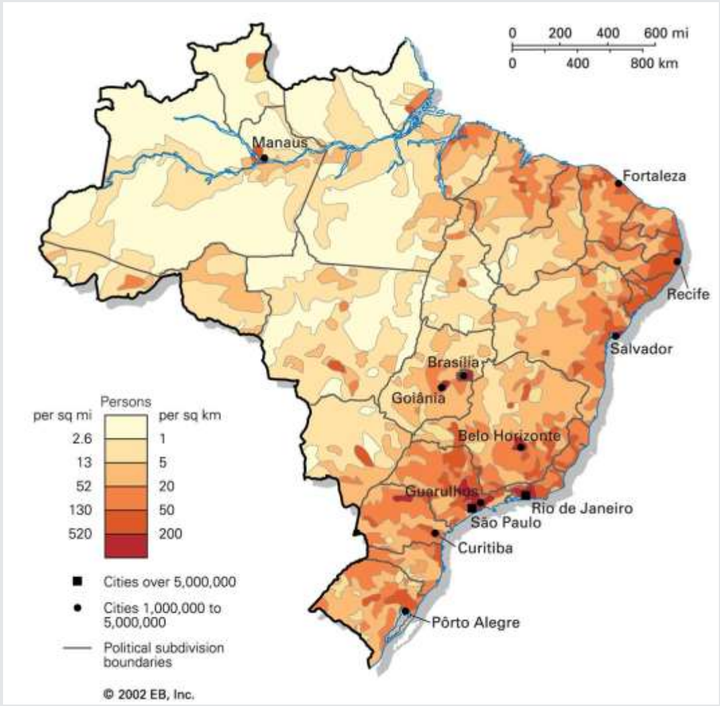
Data Exploring





Data Exploring





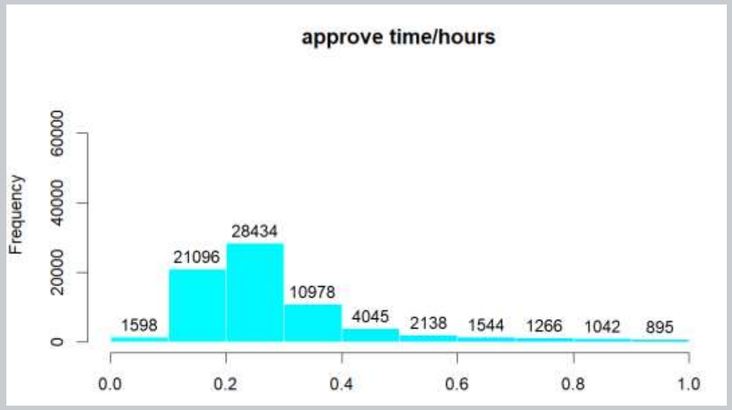
https://zh.brazilmap360.com/pdf/%E5%B7%B4%E8%A5%BF%E4%BA%BA%E5%8F%A3%E5%AF%86%E5%BA%A6%E5%9B%BEpdf.pdf

order time

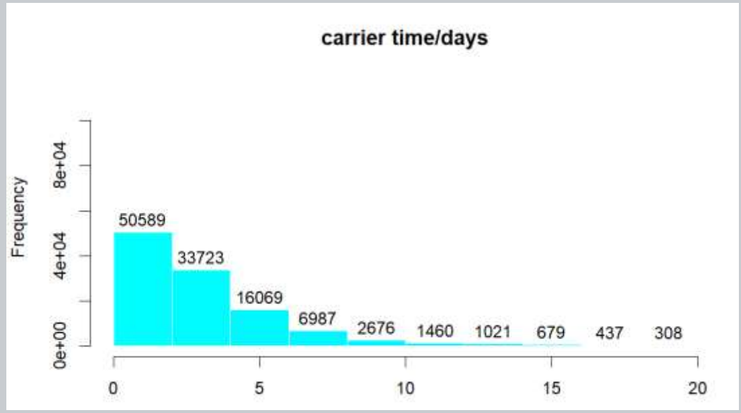
estimate time

purchase → approve → carrier → customer

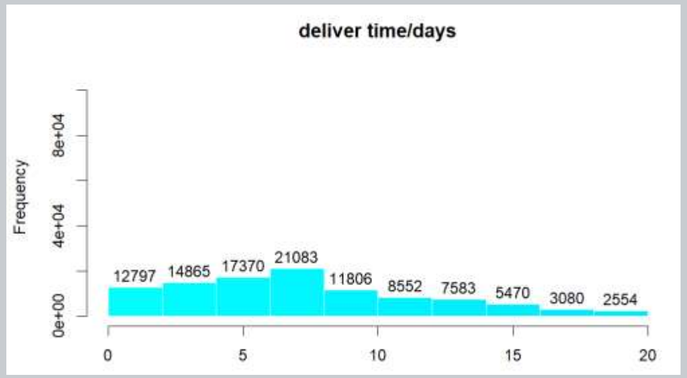
approve



carrier time



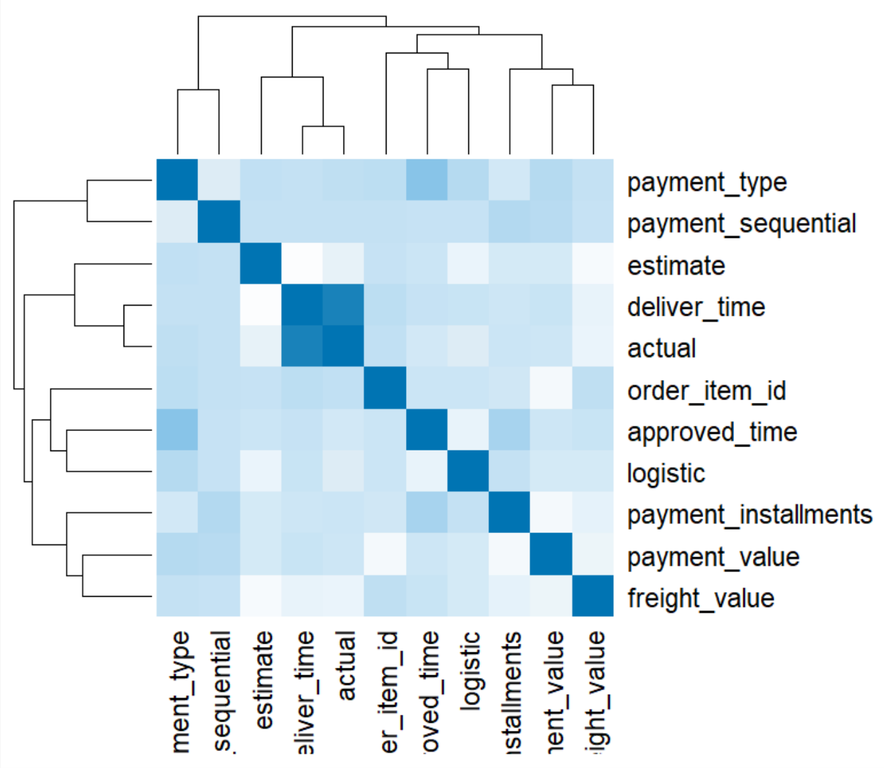
deliver time



How about the order time ?

time

estimate time?
...



payment

payment value?
freight value?
payment_sequential?
...

Let’s check the linear model!

GLM 过程

读取的观测数	114764
使用的观测数	114764

GLM 过程

因变量: actual

源	自由度	平方和	均方	F 值	Pr > F
模型	5	472416.23	94483.25	1107.37	<.0001
误差	114758	9791422.17	85.32		
校正合计	114763	10263838.39			

R 方	变异系数	均方根误差	actual 均值
0.046027	73.92610	9.237010	12.49492

源	自由度	I 型 SS	均方	F 值	Pr > F
order_item_id	1	2445.3175	2445.3175	28.66	<.0001
payment_installments	1	20936.1198	20936.1198	245.38	<.0001
payment_sequential	1	457.8356	457.8356	5.37	0.0205
payment_value	1	34085.4517	34085.4517	399.49	<.0001
freight_value	1	414491.5007	414491.5007	4857.95	<.0001

源	自由度	III 型 SS	均方	F 值	Pr > F
order_item_id	1	110.2169	110.2169	1.29	0.2557
payment_installments	1	820.5536	820.5536	9.62	0.0019
payment_sequential	1	2.2727	2.2727	0.03	0.8704
payment_value	1	4116.7647	4116.7647	48.25	<.0001
freight_value	1	414491.5007	414491.5007	4857.95	<.0001

参数	估计	标准误差	t 值	Pr > |t|
截距	9.961476193	0.08205242	121.40	<.0001
order_item_id	-0.046507299	0.04091932	-1.14	0.2557
payment_installments	0.031964446	0.01030730	3.10	0.0019
payment_sequential	-0.006537655	0.04005716	-0.16	0.8704
payment_value	-0.000823030	0.00011849	-6.95	<.0001
freight_value	0.132225223	0.00189709	69.70	<.0001

R 方	变异系数	均方根误差	actual 均值
0.046027	73.92610	9.237010	12.49492

Before this....

Outlier

16460 observations
16460/114764

check and test

Linearity
Heteroscedasticity
Normality
Collinearity

correction

Weighted Least Squars Estimation
Box-Cox
Deletion

Before this....

Linearity

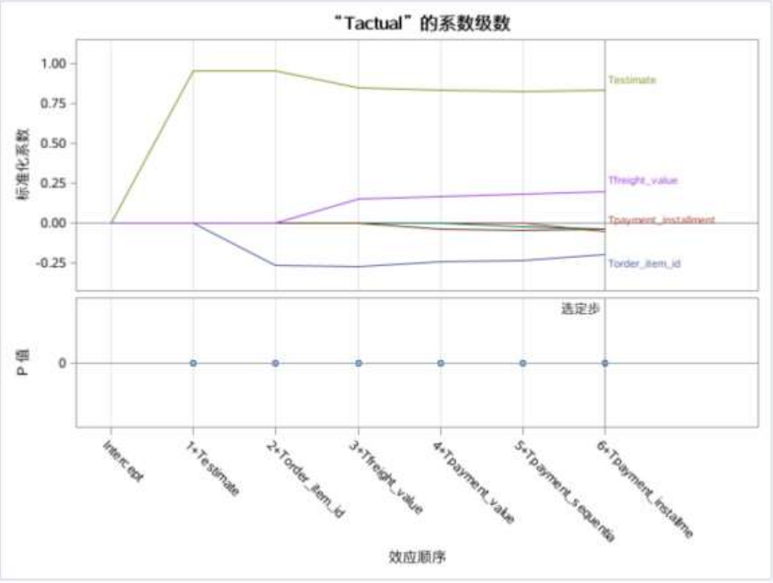
Heteroscedasticity

Normality

异方差性检验					
方程	检验	统计量	自由度	Pr > 卡方	变量
actual	White 检验	3656	27	<.0001	所有变量的叉积
	Breusch-Pagan	3193	6	<.0001	1, order_item_id, payment_ payment_value, freight_valu

正态性检验			
方程	检验统计量	值	概率
resid	Kolmogorov-Smirnov	0.10	0.0010
系统	Mardia 偏度	25822	<.0001
	Mardia 峰度	135.2	<.0001
	Henze-Zirkler T	1514	<.0001

stepwise selection



逐步选择汇总

步	进入的效应	删除的效应	引入效应数	调整R 方	AIC	CP	SBC	F 值	Pr > F
0	Intercept		1	0.0000	1226055.38	4164899340	1120185.95	0.00	1.0000
1	Testimate		2	0.9167	962903.46	346783029	857043.60	1165339	<.0001
2	Torder_item_id		3	0.9869	767020.17	54431342	661169.88	567547	<.0001
3	Tfreight_value		4	0.9994	448222.96	2579585	342382.24	2044236	<.0001
4	Tpayment_value		5	0.9995	421838.19	1987196	316007.04	29964.5	<.0001
5	Tpayment_sequential		6	0.9997	382293.41	1334811	276471.83	47942.0	<.0001
6	Tpayment_installment		7	1.0000*	105887.01*	7*	75.00*	1334806	<.0001

* 准则的最佳值

Final model

均方根误差	11.29631	R 方	0.9968
因变量均值	3.08442	调整 R 方	0.9968
变异系数	366.23755		

参数估计

变量	标签	自由度	参数估计	标准误差	t 值	Pr > |t|	容差
Intercept	Intercept	1	1.53804	0.00037446	4107.32	<.0001	.
Tpayment_installments	payment_installments Transformation	1	-0.03531	0.00008840	-399.50	<.0001	0.20393
Tpayment_sequential	payment_sequential Transformation	1	-0.10689	0.00032155	-332.42	<.0001	0.42694
Tpayment_value	payment_value Transformation	1	-0.00069478	0.00000188	-369.93	<.0001	0.19057
Tfreight_value	freight_value Transformation	1	0.02522	0.00002971	848.76	<.0001	0.28380
Testimate	estimate Transformation	1	0.05937	0.00002865	2071.93	<.0001	0.21375

04

Sentiment Analysis

Sentiment Polarity & Subjectivity

Two Features of the olist_order_review_dataset

- Both the comment title and comment message are included, the latter one contains more information.
- Consists of Portuguese and emojis, can't directly use tools like VADER. LeIA is used instead.

review_comment_title	review_comment_message	tb_title_Pol	tb_title_Subj	tb_message_Pol	tb_message_Subj
recomendo	aparelho eficiente. no site a marca do aparelh...	0	0	0	0
Super recomendo	Vendedor confiável, produto ok e entrega antes...	0.333333	0.666667	0.5	0.5
Não chegou meu produto	Péssimo	0	0	0	0

title_compound	title_neg	title_neu	title_pos	message_compound	message_neg	message_neu	message_pos
0.3612	0	0	1	0.4215	0	0.896	0.104
0.7506	0	0	1	0.6705	0	0.522	0.478
-0.296	0.423	0.577	0	0	0	1	0
0.3612	0	0	1	0	0	1	0
0.4215	0	0.263	0.737	0.1901	0.062	0.831	0.107

Sentiment Polarity & Subjectivity

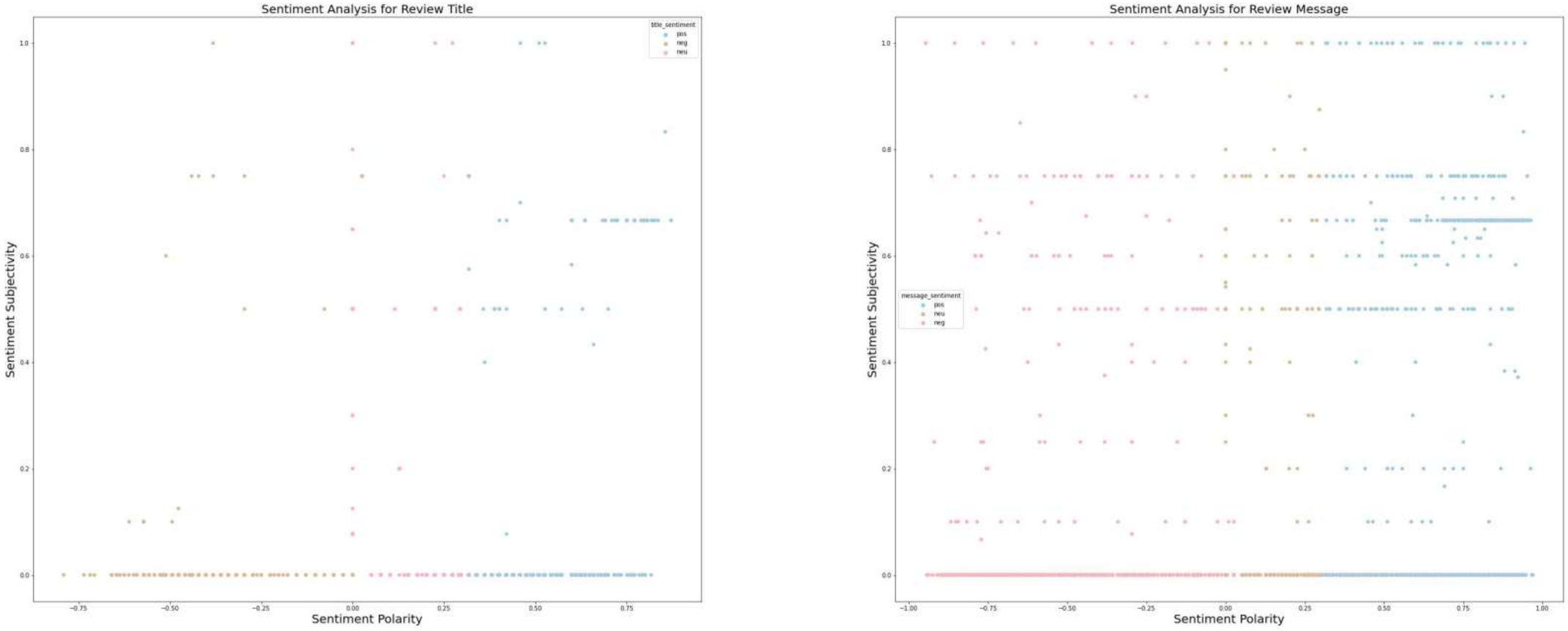
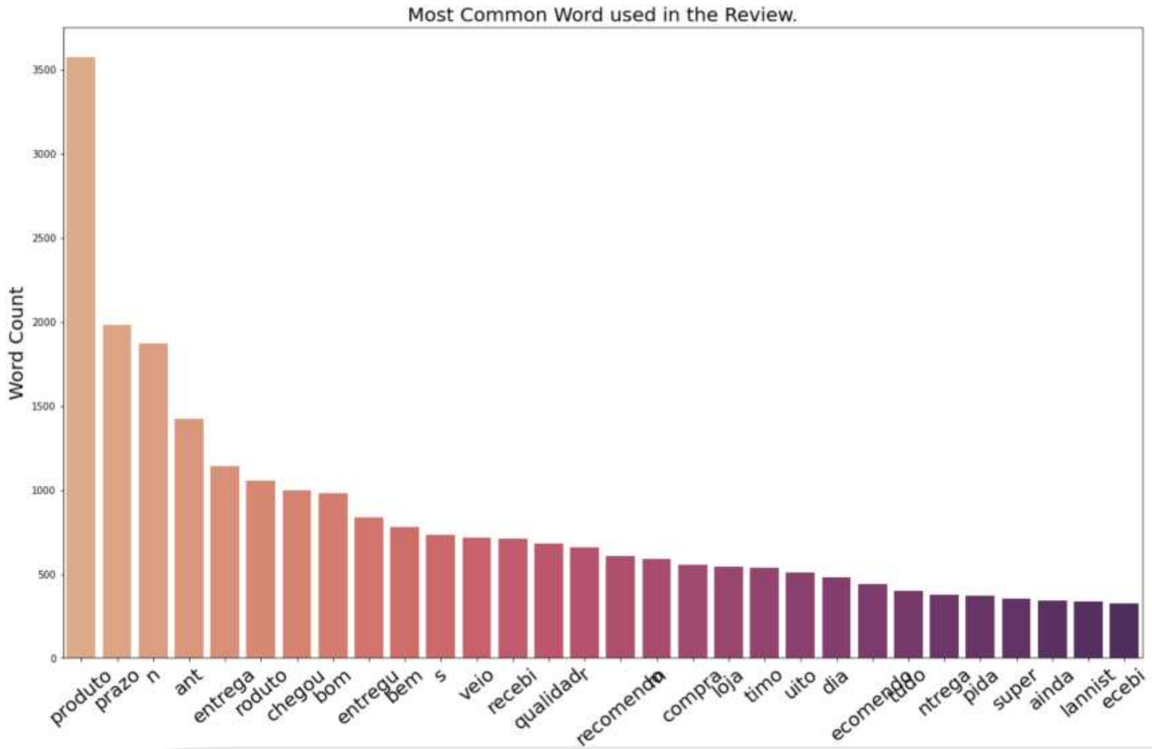


Figure 1: Polarity vs. Subjectivity for title(left) and Polarity vs. Subjectivity for message(right).

Wordcloud for Sentiment Polarity



- Preprocessing;
- nltk: STOPWORDS, PorterStemmer
- Customers care a lot about the **transportation and quality** of the product
- '**Recomendo**': customers who left comments are generally satisfied with the product they bought.

Wordcloud for Sentiment Polarity

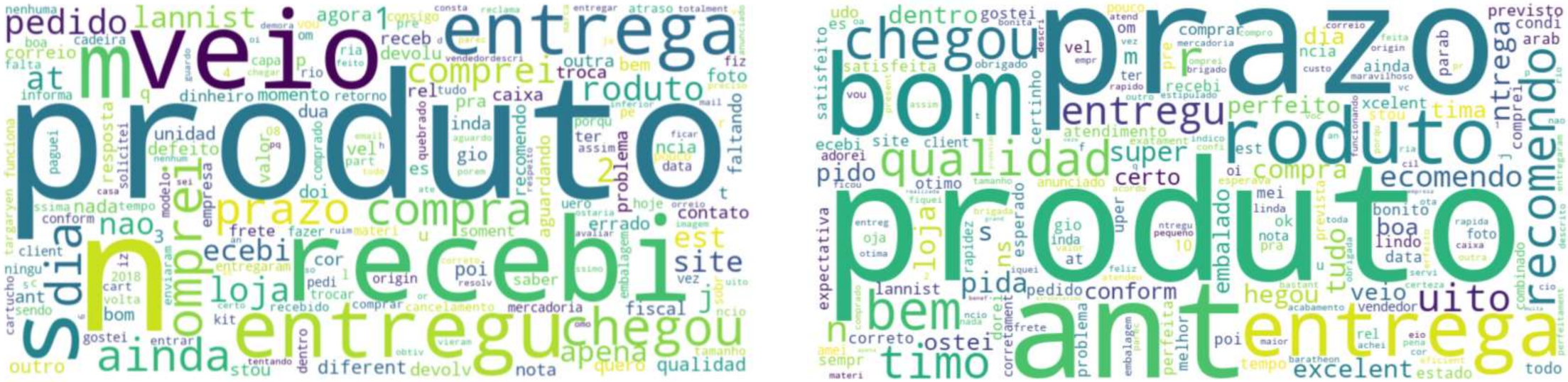
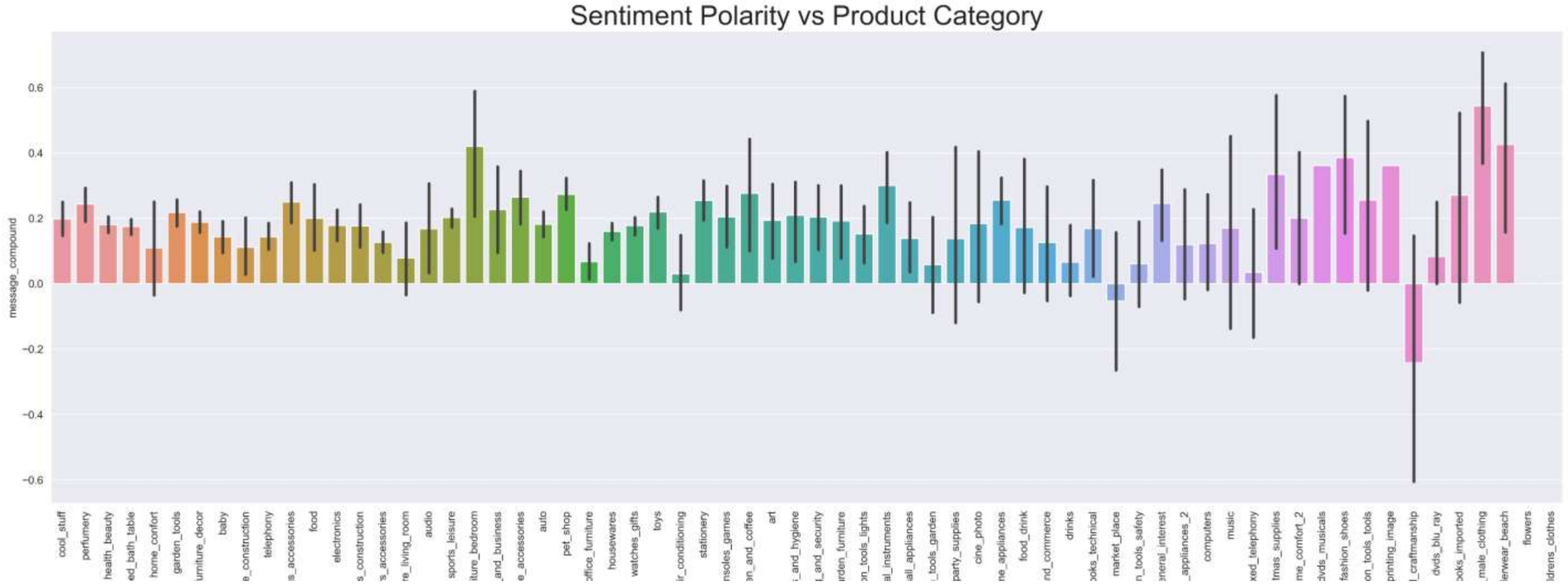


Figure 3: Wordcloud for review of customers who are not satisfied(left) and who are satisfied(right).

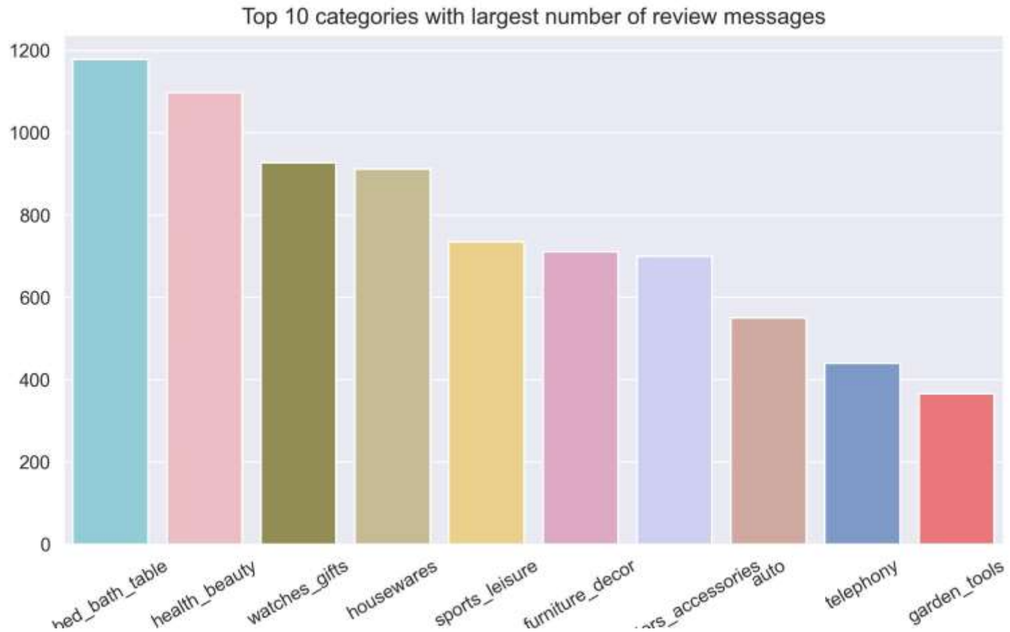
- Satisfaction: review score is between 1-3 means not satisfied
- Reviews concerned with transportation time are from unsatisfied customers: shed light on improvements for sellers

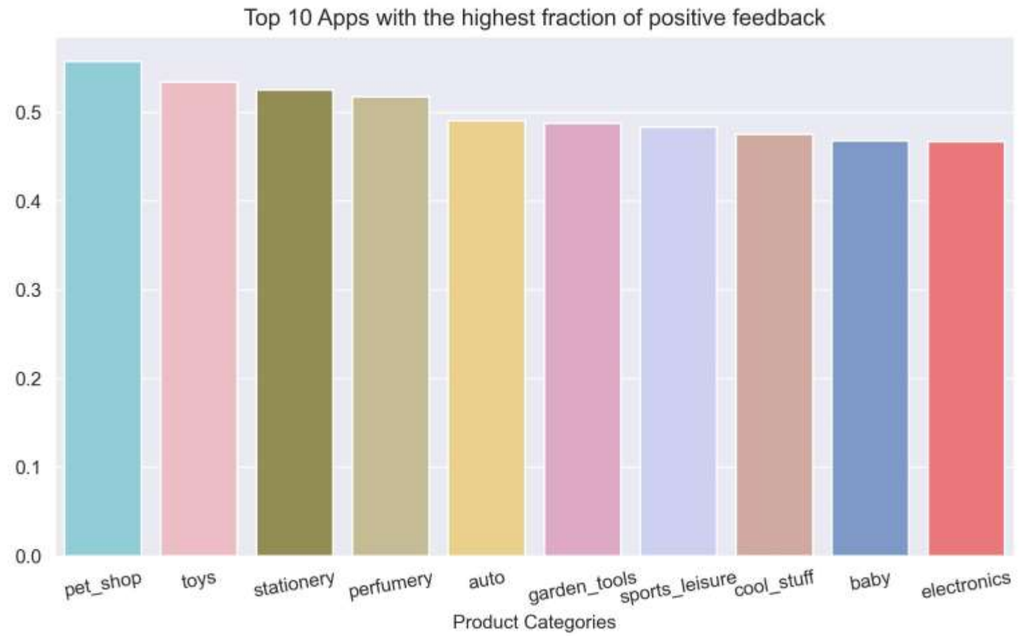
Which Products Receive More Positive Reviews?



- Categories which averagely receive the top 3 highest positive review sentiment: **fashion male clothing, fasion underware beach** and **furniture bedroom**
- Noticeable: **art and craftmanship** receive an averagely negative review

Which Products Receive More Positive Reviews?





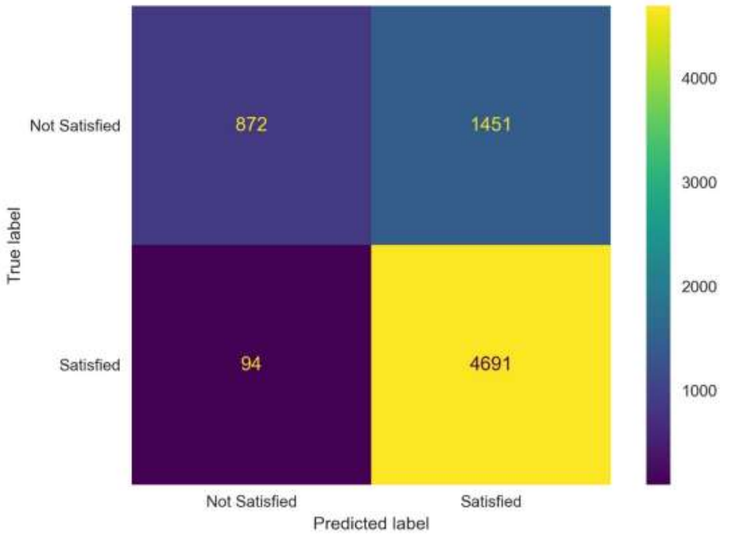
- **Bed bath table** and **healthy beauty** receive the top 2 largest number of reviews according to the product category
- **Pet shop** and **toys** receive the top 2 highest fraction of positive reviews

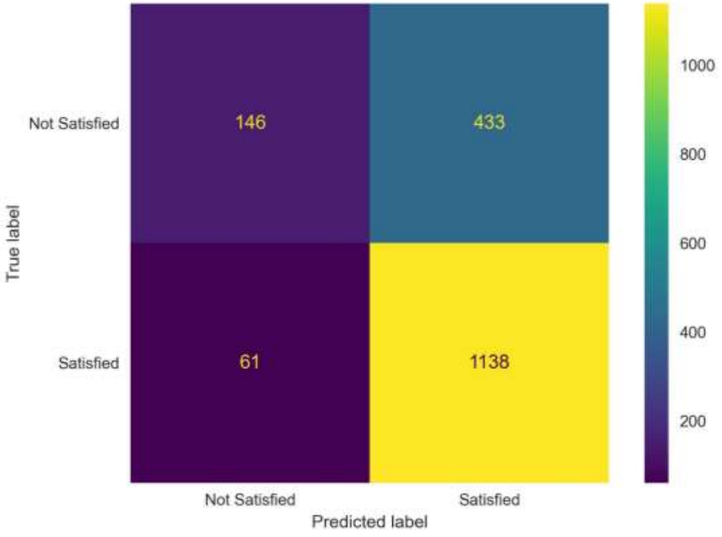
Geographical Analysis with Sentiment

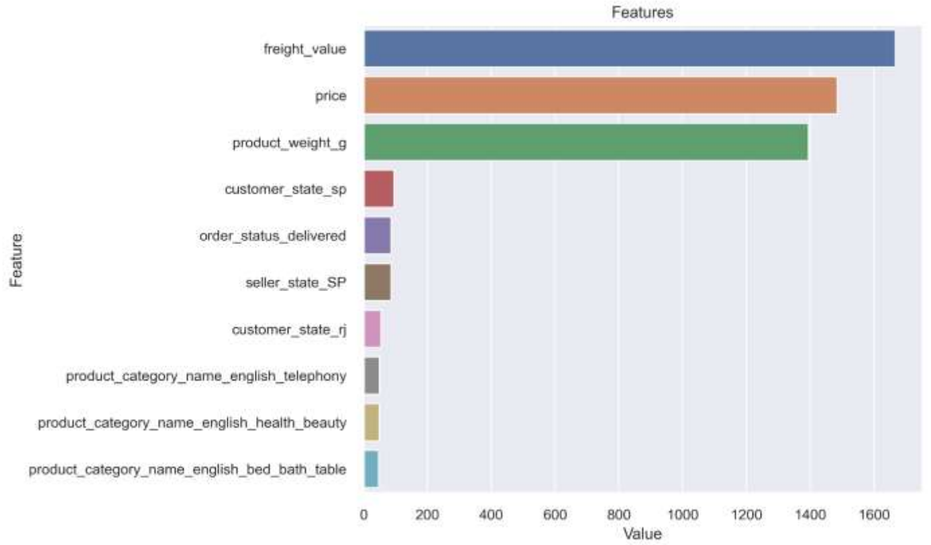
Figure 5: Geographical visualization of sentiment according to the customer's state/city(left) and the seller's state/city(right).

- Top 3 state which receive the highest average sentiment polarity are **ms**(0.599400), **pi**(0.522400), **pe**(0.243585). Top 3 state which give the highest average sentiment polarity are **rr**(0.482938), **ap**(0.305560), **am**(0.273921).
- Top 3 cities which receive the highest average sentiment polarity are **nova lima**(0.93710), **irati**(0.89340), **balneario camboriu**(0.88005). Top 3 cities which give the highest average sentiment polarity are **aparecida**(0.9286), **sitio novo**(0.9214), **andre da rocha**(0.9081).

LightGBM for Customer Satisfaction







- Fitted using **freight value, price, product weight g, order status delivered, product category name english, customer state and seller state**.
- **Preprocessing**: one hot encoding, standardized the numerical variables, and deleted the outliers according to quantile.
- Random forest: variable importance.

Thank you
for your watching