AliAntTest

笔记本:	Temp		
创建时间:	2020/4/18 15:28	**更新时间:**	2020/4/18 21:46
作者:	eric_ren@aliyun.com		

Summary

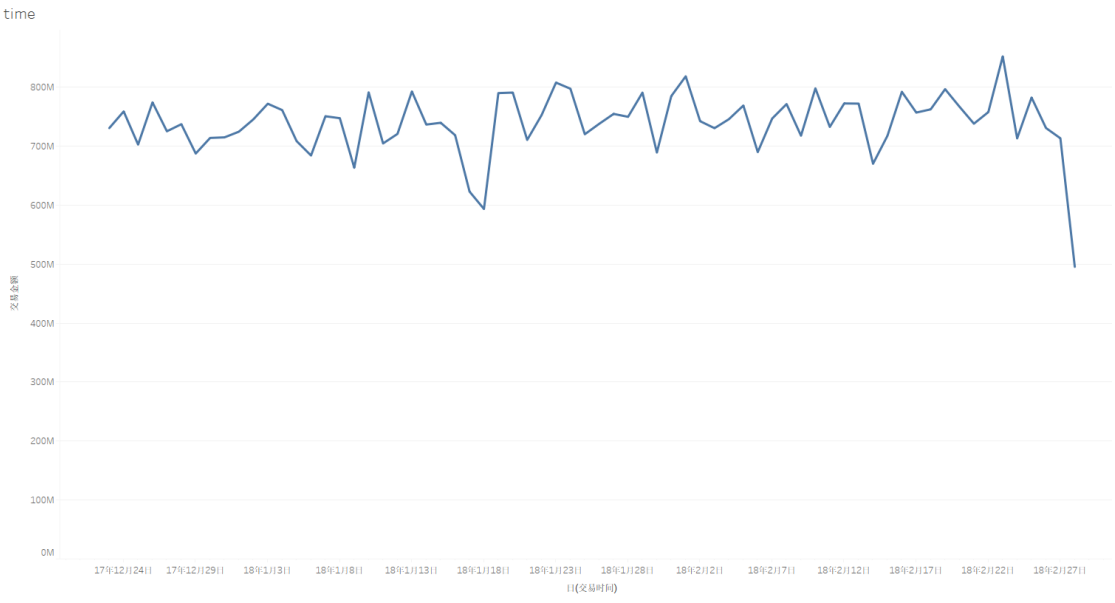
这里按照下面步骤来探索这个题目：

- 数据探索：为后续特征提取提供数据支撑；
- 模型选择：用于检测异常点的方法比较多，对不同算法进行简单比较；
- 特征提取&模型构建。

数据预览

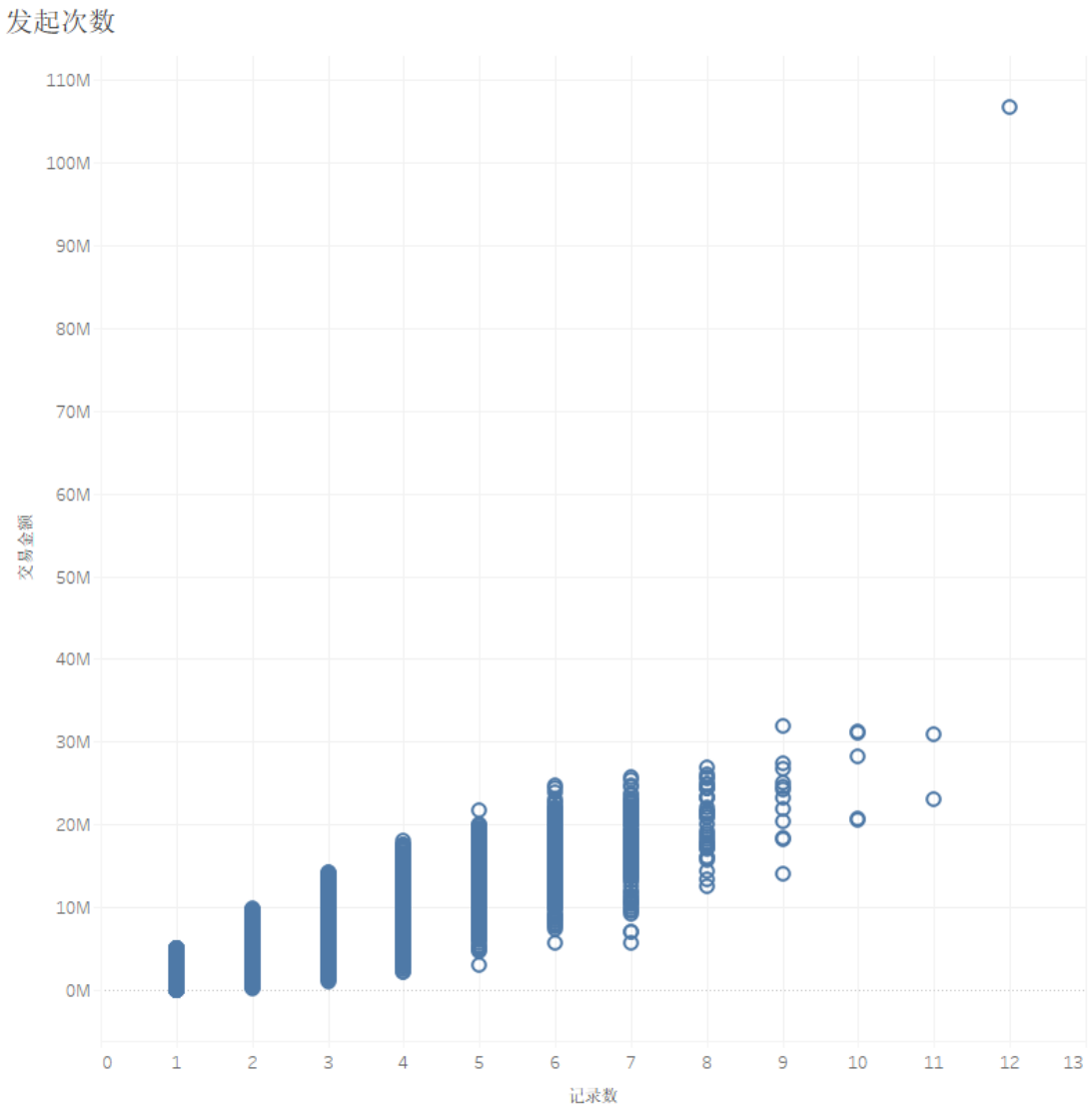
时间维度

- 看不出异常：



发起方

- 交易金额有离群点：



- 查看数据：

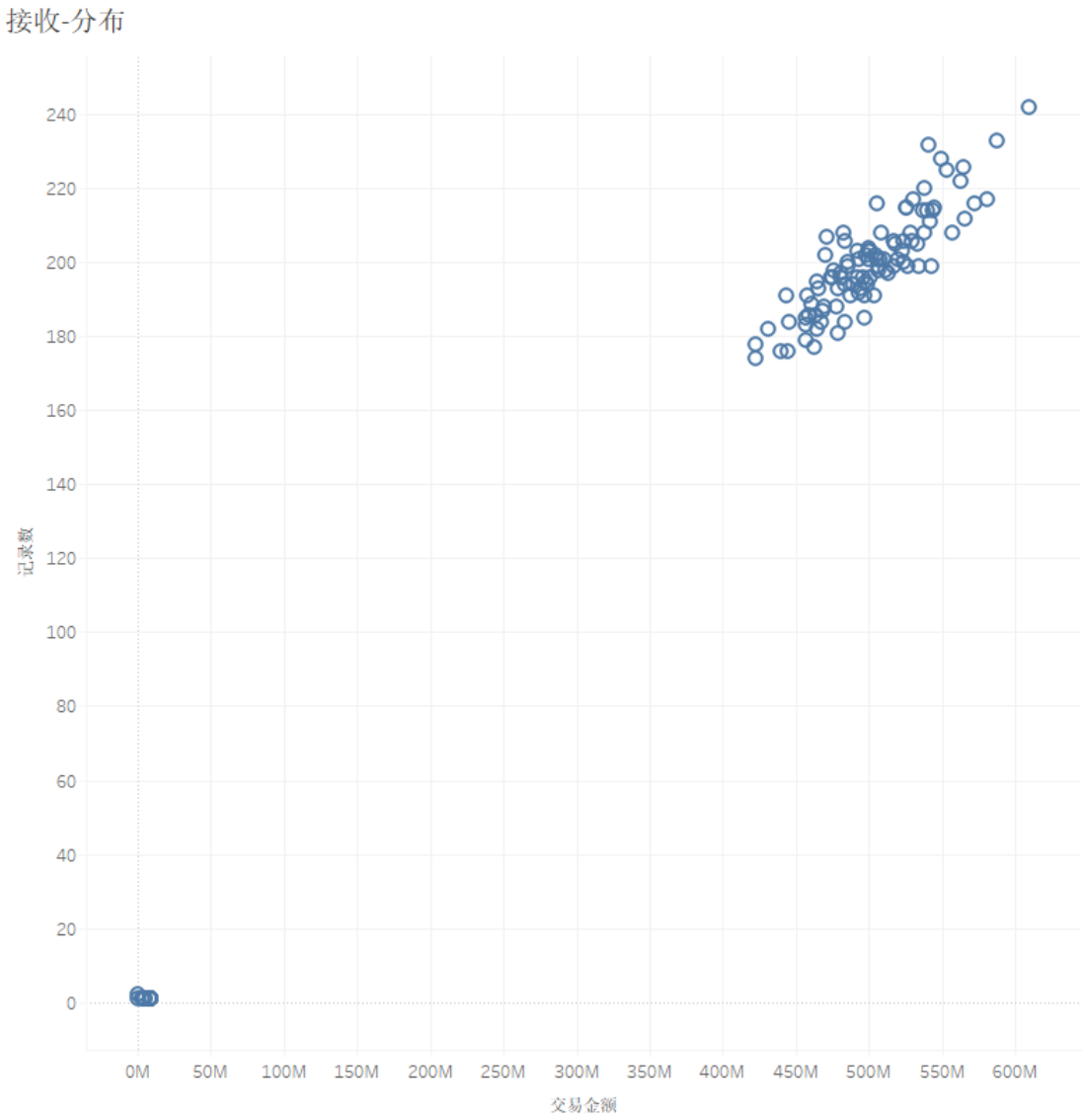
交易时间	交易金额(数据精)	发起方id	发起方年龄	发起方所处地区	接收方ID	通道	转账附言	交易金额	每笔价格
2018/2/1 13:18:25	8,863,830	1137	59	福建	1434	跨境转账	转账	8,888,800	8,888,800.00
2018/2/1 13:18:25	8,863,830	1137	59	福建	658	跨境转账	转账	8,888,800	8,888,800.00
2018/2/2 13:18:25	8,863,830	1137	59	福建	1235	跨境转账	转账	8,888,800	8,888,800.00
2018/2/3 13:18:25	8,863,830	1137	59	福建	195	跨境转账	转账	8,888,800	8,888,800.00
2018/2/4 13:18:25	8,863,830	1137	59	福建	295	跨境转账	转账	8,888,800	8,888,800.00
2018/2/5 13:18:25	8,863,830	1137	59	福建	335	跨境转账	转账	8,888,800	8,888,800.00
2018/2/6 13:18:25	8,863,830	1137	59	福建	952	跨境转账	转账	8,888,800	8,888,800.00
2018/2/7 13:18:25	8,863,830	1137	59	福建	94	跨境转账	转账	8,888,800	8,888,800.00
2018/2/8 13:18:25	8,863,830	1137	59	福建	832	跨境转账	转账	8,888,800	8,888,800.00
2018/2/9 13:18:25	8,863,830	1137	59	福建	541	跨境转账	转账	8,888,800	8,888,800.00
2018/2/10 13:18:25	8,863,830	1137	59	福建	725	跨境转账	转账	8,888,800	8,888,800.00
2018/2/11 13:18:25	8,863,830	1137	59	福建	435	跨境转账	转账	8,888,800	8,888,800.00

交易时间上是2018.2.1——2018.2.11每天固定时间交易的，发起ID是1137，金额都是8,888,800。

先记录下来，特征提取阶段需要考虑。

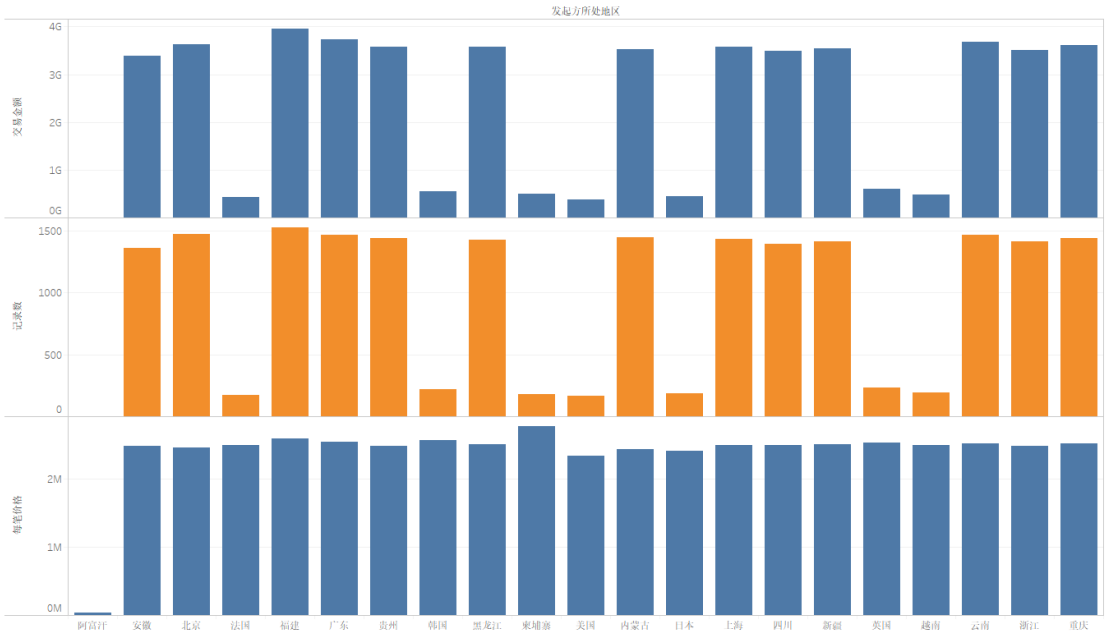
接收方

从数据上看也有离群点，特征提取阶段需要考虑：



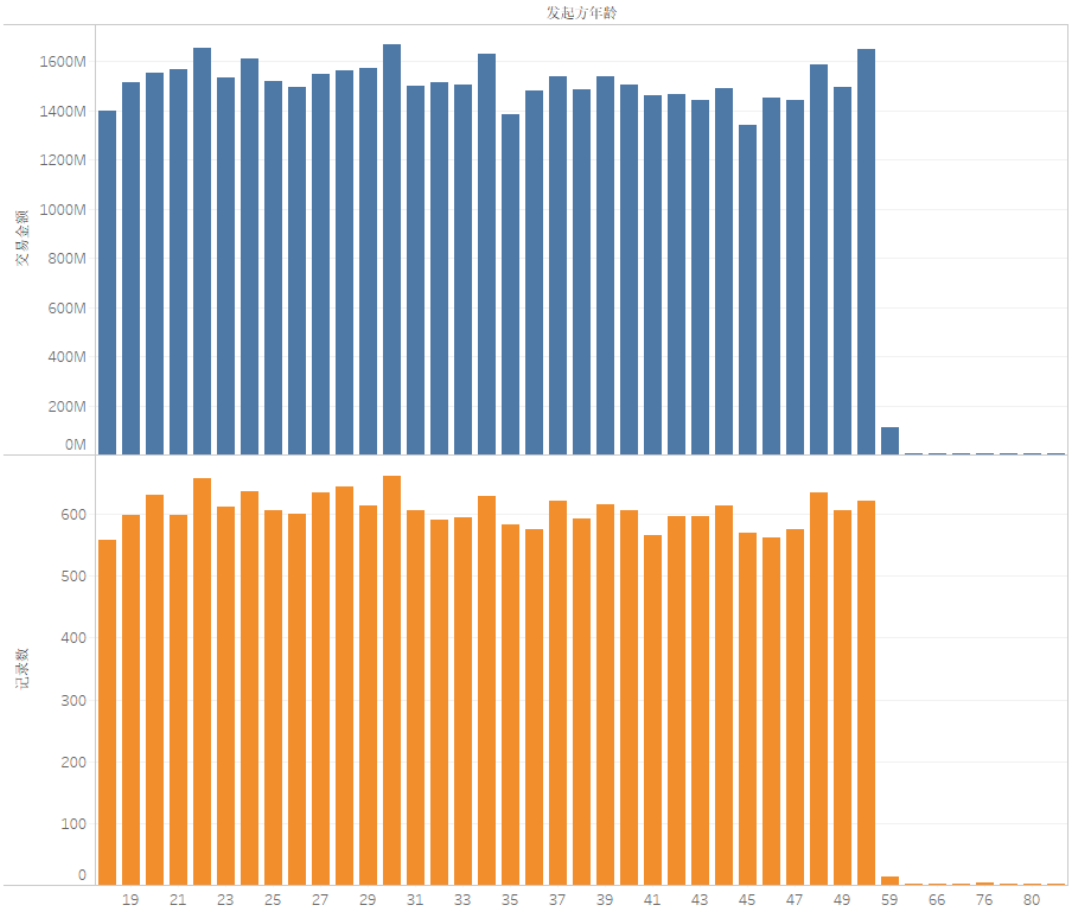
发起地区

发起地区

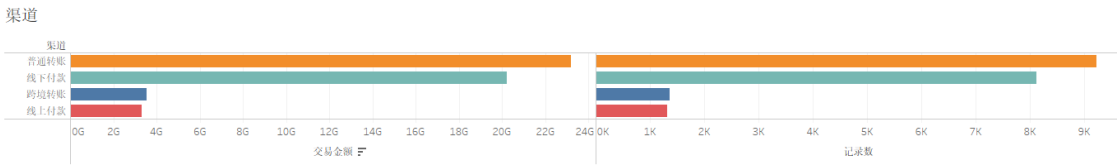


年龄分布

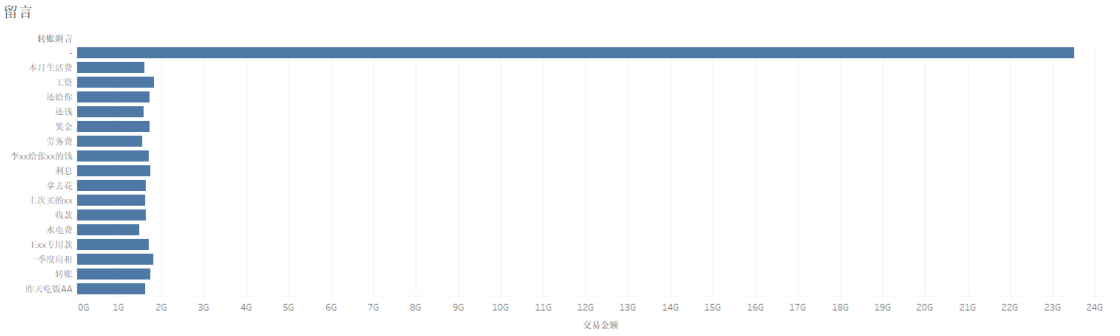
年龄



渠道统计



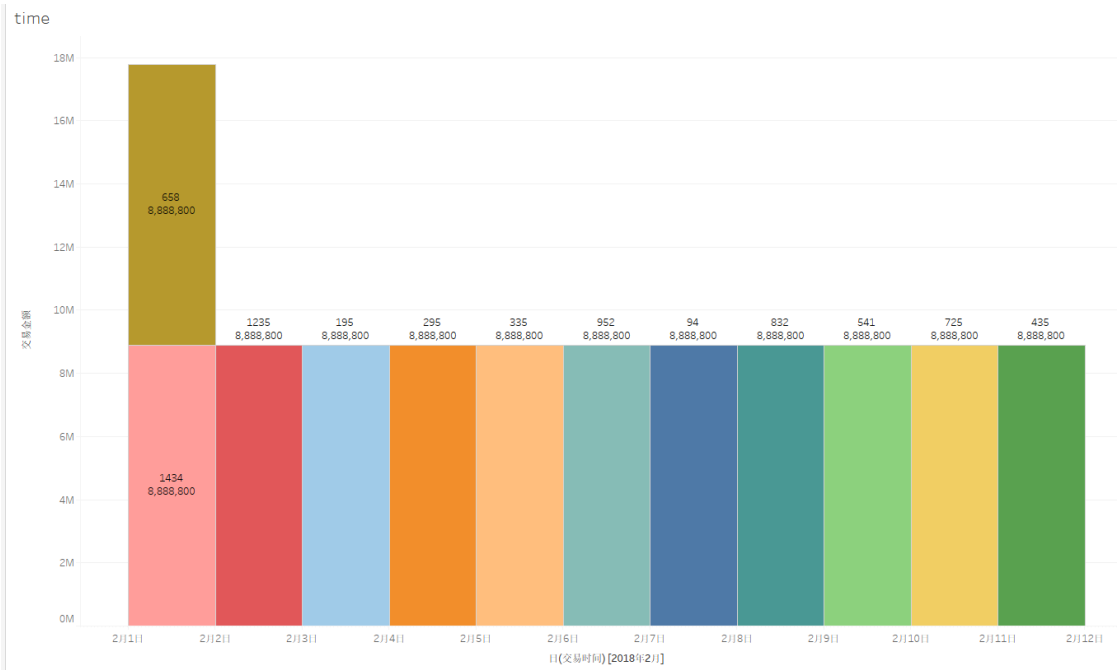
留言分布



异常分析

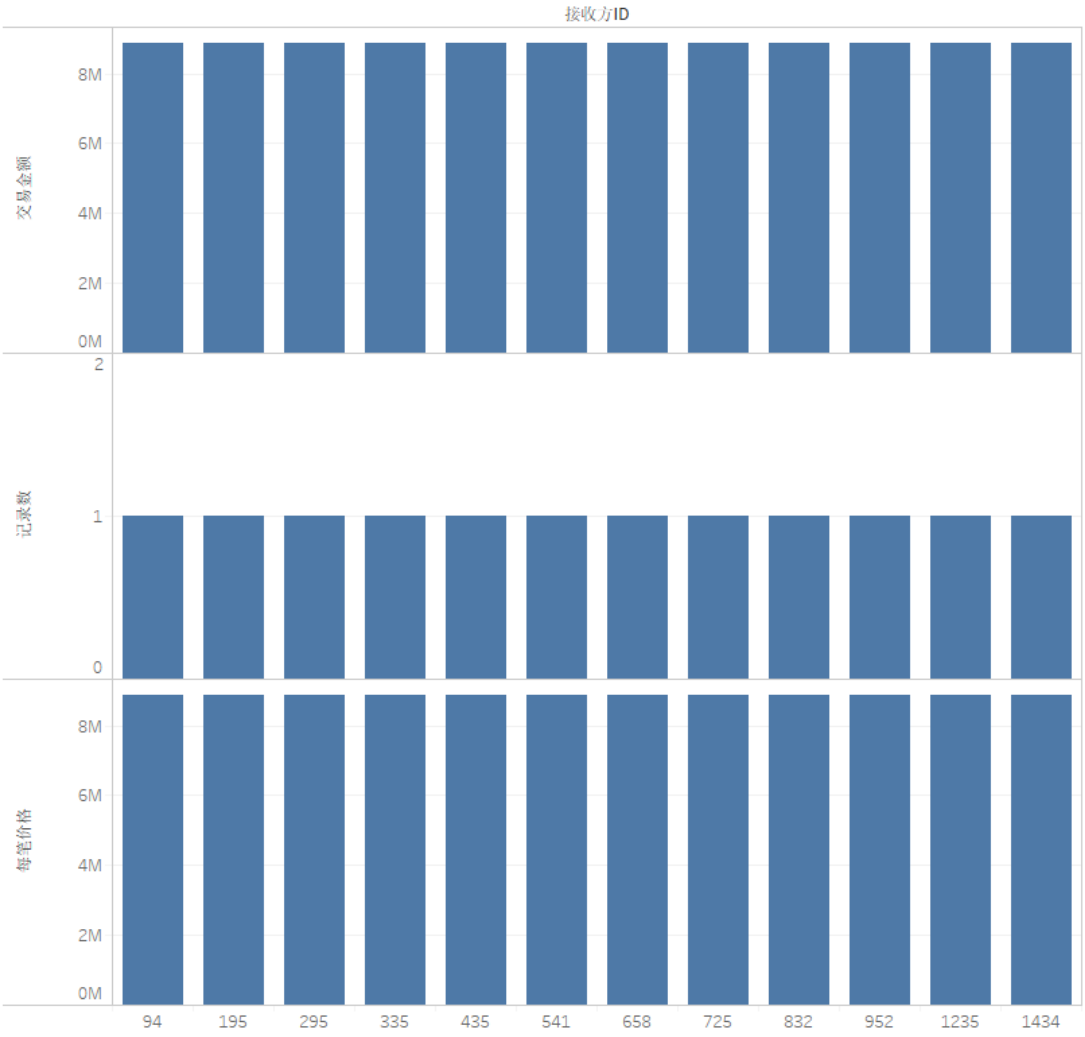
发起方1137相关的交易

- 查看1137发起交易的详细信息：



- 这些账户包括：

发起方1137的交易

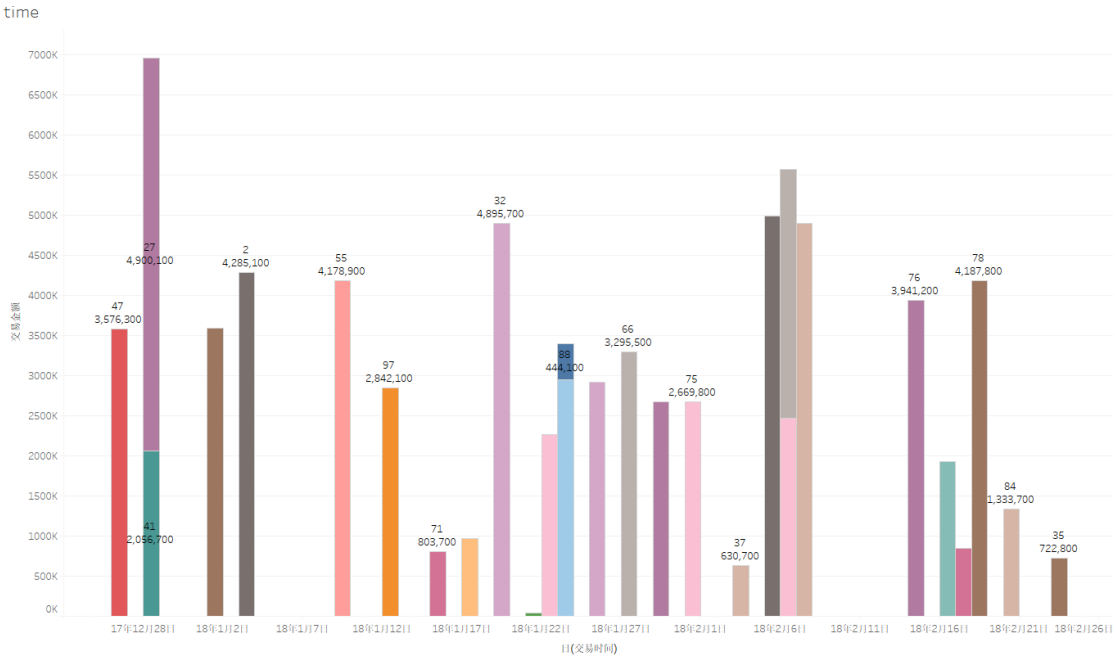


- 详细记录：

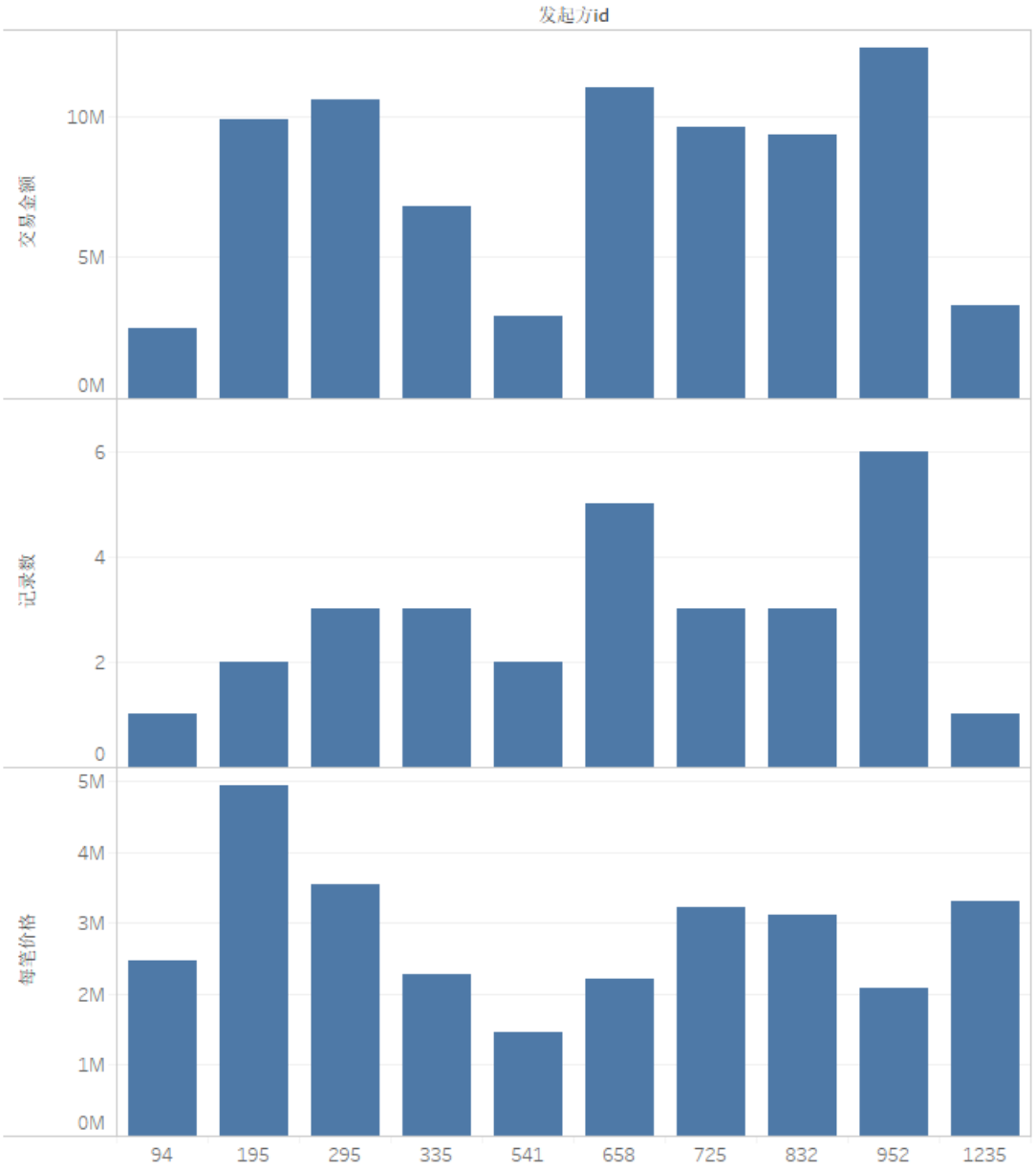
交易时间	交易金额 (数据精)	发起方id	发起方年龄	发起方所处地区	接收方ID	通道	转账附言	交易金额	每笔价格
2018/2/1 13:18:25	8,863,830	1137	59	福建	1434	跨境转账	转账	8,888,800	8,888,800.00
2018/2/1 13:18:25	8,863,830	1137	59	福建	658	跨境转账	转账	8,888,800	8,888,800.00
2018/2/2 13:18:25	8,863,830	1137	59	福建	1235	跨境转账	转账	8,888,800	8,888,800.00
2018/2/3 13:18:25	8,863,830	1137	59	福建	195	跨境转账	转账	8,888,800	8,888,800.00
2018/2/4 13:18:25	8,863,830	1137	59	福建	295	跨境转账	转账	8,888,800	8,888,800.00
2018/2/5 13:18:25	8,863,830	1137	59	福建	335	跨境转账	转账	8,888,800	8,888,800.00
2018/2/6 13:18:25	8,863,830	1137	59	福建	952	跨境转账	转账	8,888,800	8,888,800.00
2018/2/7 13:18:25	8,863,830	1137	59	福建	94	跨境转账	转账	8,888,800	8,888,800.00
2018/2/8 13:18:25	8,863,830	1137	59	福建	832	跨境转账	转账	8,888,800	8,888,800.00
2018/2/9 13:18:25	8,863,830	1137	59	福建	541	跨境转账	转账	8,888,800	8,888,800.00
2018/2/10 13:18:25	8,863,830	1137	59	福建	725	跨境转账	转账	8,888,800	8,888,800.00
2018/2/11 13:18:25	8,863,830	1137	59	福建	435	跨境转账	转账	8,888,800	8,888,800.00

从1137接收到的金额流到了哪里?

- 以从1137接收财富的帐号做数据集，探索交易信息：


time
7000K
6500K
6000K
5500K
5000K
4500K
4000K
3500K
3000K
2500K
2000K
1500K
1000K
500K
0K
交易金额
47
3,576,300
27
4,900,100
41
2,056,700
2
4,285,100
55
4,178,900
97
2,842,100
71
803,700
32
4,895,700
88
444,100
66
3,295,500
75
2,669,800
37
630,700
76
3,941,200
78
4,187,800
84
1,333,700
35
722,800
17年12月28日
18年1月2日
18年1月7日
18年1月12日
18年1月17日
18年1月22日
18年1月27日
18年2月1日
18年2月6日
18年2月11日
18年2月16日
18年2月21日
18年2月26日
日(交易时间)

与1137关联的账户的财富流向



未发现明显异常。

留言有些异常

有很多特定格式的留言，比如：李xx给张xx的钱、王xx专用款。

交易时间	交易金额(数据桶)	发起方id	发起方年龄	发起方所处地区	接收方ID	渠道	转账附言	交易金额	
2017/12/23 1:33:25	3,002,265	6999	29	福建	56	普通转账	李xx给张xx的钱	3,066,200	3,06
2017/12/23 2:33:47	285,930	4656	49	广东	66	普通转账	李xx给张xx的钱	287,600	28
2017/12/23 2:56:58	1,429,650	1245	28	浙江	96	普通转账	李xx给张xx的钱	1,519,700	1,51
2017/12/23 4:36:08	1,572,615	4570	38	黑龙江	87	普通转账	李xx给张xx的钱	1,706,500	1,70
2017/12/23 6:25:09	4,574,880	6595	34	重庆	5	普通转账	李xx给张xx的钱	4,648,100	4,64
2017/12/23 9:52:17	4,860,810	4718	30	法国	92	跨境转账	李xx给张xx的钱	4,909,200	4,90
2017/12/23 13:42:02	1,858,545	52	24	广东	4	普通转账	李xx给张xx的钱	1,881,600	1,88
2017/12/23 17:40:17	1,858,545	609	25	云南	11	普通转账	李xx给张xx的钱	1,920,400	1,92
2017/12/23 19:08:17	4,145,985	3916	25	内蒙古	98	普通转账	李xx给张xx的钱	4,182,000	4,18
2017/12/23 19:23:21	1,000,755	6541	26	重庆	86	普通转账	李xx给张xx的钱	1,054,500	1,05
2017/12/23 21:53:19	2,716,335	4822	40	北京	92	普通转账	李xx给张xx的钱	2,838,800	2,83
2017/12/23 21:56:13	4,003,020	4013	46	安徽	40	普通转账	李xx给张xx的钱	4,049,200	4,04
2017/12/23 23:00:25	714,825	3786	31	新疆	60	普通转账	李xx给张xx的钱	758,800	75
2017/12/24 3:47:01	428,895	4767	32	云南	87	普通转账	李xx给张xx的钱	447,200	44
2017/12/24 5:27:49	142,965	655	46	日本	73	跨境转账	李xx给张xx的钱	200,800	20
2017/12/24 11:29:28	2,430,405	3589	26	广东	52	普通转账	李xx给张xx的钱	2,546,600	2,54

交易时间	交易金额(数据桶)	发起方id	发起方年龄	发起方所处地区	接收方ID	渠道	转账附言	交易金额	每笔
2017/12/23 2:40:12	1,286,685	6020	36	福建	57	普通转账	王xx专用款	1,344,000	1,344,00
2017/12/23 3:15:49	2,859,300	2967	48	新疆	56	普通转账	王xx专用款	2,988,500	2,988,50
2017/12/23 3:28:13	4,145,985	2472	26	新疆	22	普通转账	王xx专用款	4,262,300	4,262,30
2017/12/23 5:48:22	4,003,020	4088	19	四川	91	普通转账	王xx专用款	4,033,500	4,033,50
2017/12/23 6:00:58	2,001,510	2778	33	法国	80	跨境转账	王xx专用款	2,007,900	2,007,90
2017/12/23 7:50:00	2,001,510	3061	25	浙江	20	普通转账	王xx专用款	2,053,000	2,053,00
2017/12/23 7:54:42	2,287,440	5682	27	内蒙古	20	普通转账	王xx专用款	2,336,500	2,336,50
2017/12/23 7:58:23	1,572,615	3903	42	广东	98	普通转账	王xx专用款	1,587,900	1,587,90
2017/12/23 10:22:05	2,287,440	5667	21	云南	98	普通转账	王xx专用款	2,414,700	2,414,70
2017/12/23 12:44:42	0	1156	37	韩国	32	跨境转账	王xx专用款	133,700	133,70
2017/12/23 14:11:50	285,930	153	23	北京	39	普通转账	王xx专用款	306,700	306,70
2017/12/23 17:38:31	1,858,545	4800	19	北京	48	普通转账	王xx专用款	1,938,600	1,938,60
2017/12/23 18:37:37	3,431,160	6692	40	浙江	62	普通转账	王xx专用款	3,512,300	3,512,30
2017/12/23 19:04:45	714,825	6789	23	黑龙江	11	普通转账	王xx专用款	766,900	766,90
2017/12/23 19:47:48	3,717,090	6305	38	安徽	32	普通转账	王xx专用款	3,778,300	3,778,30
2017/12/24 0:38:15	1,000,755	1856	35	浙江	17	普通转账	王xx专用款	1,007,800	1,007,80

那这个是否需要分析？

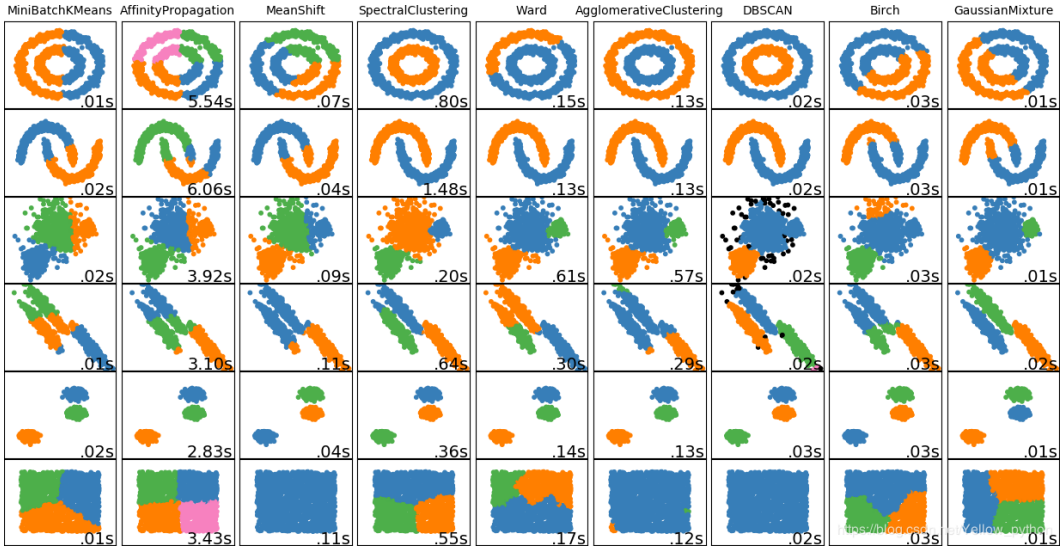
从我个人的理解，留言栏可能会用来传递暗语。但考虑到现实中留言种类繁多，大部分都是噪音，且暗语类型的信息无特定格式，属于后验信息（也就是当你知道这是异常交易后再来看暗语可能会发现一些规律），所以无需考虑。

结论：这个数据集里重复的特定格式的留言可以理解为巧合，不予考虑。

模型的选择

风险监控问题的本质是分类问题，可选的模型有：

- 非监督学习——聚类；
  几种常用聚类方法的比较：



在聚类效果上，谱聚类、DBscan的方法是非线性的，精度最优，但DBscan要探索一个阈值，灵活性低。

而谱聚类方法可以通过指定cluster个数来聚类，风险监控场景可以看作二分类场景，聚类个数一定是2，所以谱聚类比较贴合这个场景 。

- 监督学习——深度学习模型；

  监督学习需要有大量标注数据，在这个场景下显然是不适合的。

  不过阿里应该不缺这些数据，比如用户在被盗、被骗之后大概率会联系阿里，告知哪些交易非本人操作，抑或哪些交易是被骗的，这些反馈可以比较准确的帮助阿里沉淀带标签的数据。
- 异常点检测

  这一类算法是在风控模型里比较常用的一种方法，常用的是Isolation Forest算法。

在这里为了照顾模型精度和开发工作量，这里选用两种聚类方法K-menas、谱聚类和Isolation Forest算法来构建模型。

特征提取&模型构建

这部分内容详见：异常点检测模型小测试