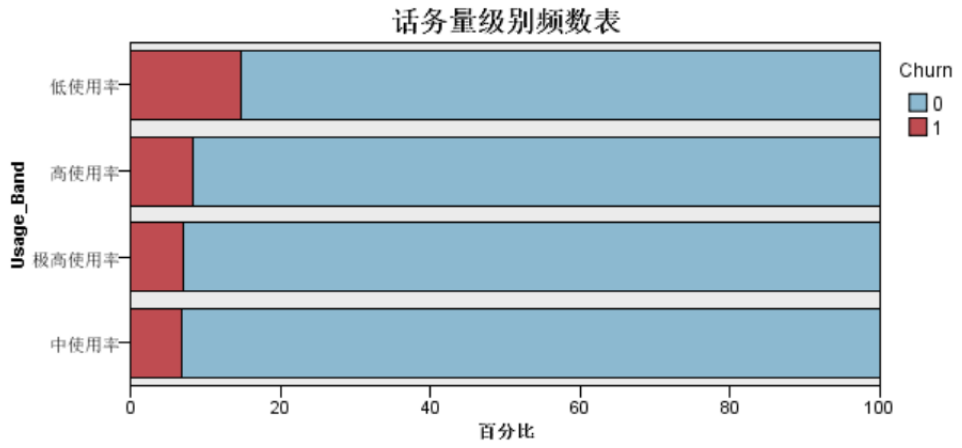
三、数据探索性分析

1、离散型变量的探索性分析方法

对无序型离散变量而言，以本案例中的手机品牌为例，对于名义型离散变量,关注的是该变量的取值分别有哪些，各个取值占比是多少。从表格上看，列出离散变量各个取值的数量和占比即可:

值	比例	%	计数
ASAD170		13.86	2571
ASAD90		3.46	641
BS110		17.37	3222
BS210		5.63	1045
CAS30		2.21	410
CAS60		2.59	481
S50		22.94	4256
S80		16.31	3025
SOP10		0.46	85
SOP20		0.56	104
WC95		14.61	2710

对有序型商散变量而言，有序型离散变量之间是可以比较大小的，因此还可以通过累积频数和崇积频率的方式来对数据进行展现。以话务量级别为例，可以做出话务量级别频数表，从表中的累积百分比可以看出，13.7%的客户属于低使用率，而 95.8%的客户在高使用率及以下。



2、连续型变量的探索性分析方法

对于连续型变量，通常可以使用描述统计量和图形两种方法来进行探索性分析。

- 使用描述统计量：对于连续型变量，常见的描述统计量包括反应变量集中趋势的

均值、中位数等；反应分散趋势的最小值、最大值、全距、标准差、变异系数等；反应分布形态的偏度和峰度。

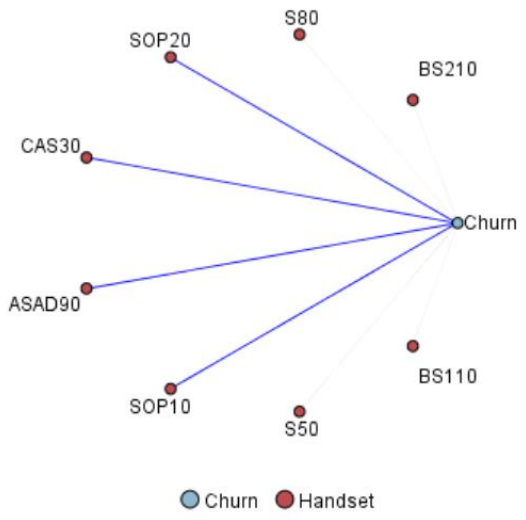
• 使用图形：对于连续型变量，主要通过直方图和箱线图的方式来对数据的分布状况进行考察。

3、变量之间关系的探索性分析方法

1）离散变量与离散变量

离散变量与离散变量之间的关系可以使用条形图进行查看，将其中一个变量在图形中 用不同的颜色显示来直观地观察出变量之间的关系，也可以使用网络图开显示，通过各个要素之间是有线条联系、线条粗线来显示是否有关系以及关系强弱。

例如，不同手机品牌的流失情况有着明显区别，在网络图中可以得到和条形图一样的结论，即 ASAD90、CAS30、SOPIO 和 SOP20 四个品牌的手机与流失关系密切。



如果希望得到两个离散变量之间关系的量化描述，可以使用交叉表来显示，从下图中可以看出每个手机品牌的流失与不流失人数及百分比，而最下面的卡方值和概率则表明从统计意义上两者是否无关，在这个案例中，看到概率=0，是一个很小的数值，说明两者的关系是显著的。

Handset

Churn		ASAD170	ASAD90	BS110	BS210	CAS30	CAS60	S50	S80	SOP10	SOP20	WC95	总计
0	计数	2556	224	2983	989	153	477	3924	2981	28	35	2698	17048
	期望值	2362.825	589.098	2961.114	960.386	376.802	442.053	3911.390	2780.065	78.118	95.579	2490.570	17048
	行百分比	14.993	1.314	17.498	5.801	0.897	2.798	23.017	17.486	0.164	0.205	15.826	100
	列百分比	99.417	34.945	92.582	94.641	37.317	99.168	92.199	98.545	32.941	33.654	99.557	91.903
	总百分比	13.779	1.208	16.081	5.332	0.825	2.571	21.154	16.070	0.151	0.189	14.544	91.903
1	计数	15	417	239	56	257	4	332	44	57	69	12	1502
	期望值	208.175	51.902	260.886	84.614	33.198	38.947	344.610	244.935	6.882	8.421	219.430	1502
	行百分比	0.999	27.763	15.912	3.728	17.111	0.266	22.104	2.929	3.795	4.594	0.799	100
	列百分比	0.583	65.055	7.418	5.359	62.683	0.832	7.801	1.455	67.059	66.346	0.443	8.097
	总百分比	0.081	2.248	1.288	0.302	1.385	0.022	1.790	0.237	0.307	0.372	0.065	8.097
总计	计数	2571	641	3222	1045	410	481	4256	3025	85	104	2710	18550
	期望值	2571	641	3222	1045	410	481	4256	3025	85	104	2710	18550
	行百分比	13.860	3.456	17.369	5.633	2.210	2.593	22.943	16.307	0.458	0.561	14.609	100
	列百分比	100	100	100	100	100	100	100	100	100	100	100	100
	总百分比	13.860	3.456	17.369	5.633	2.210	2.593	22.943	16.307	0.458	0.561	14.609	100

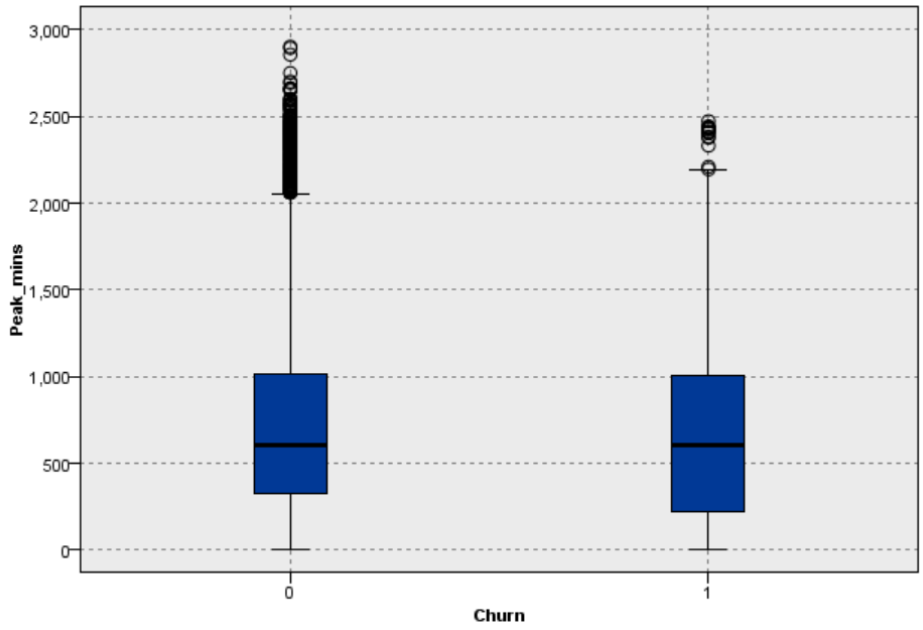
单元格内容: 字段的交叉列表 (包括缺失值)

卡方 = 5,942.412, df = 10, 概率 = 0

2) 离散变量与连续变量

对于离散变量和连续变量之间的关系，可以使用直方图进行查看，将其中的离散变量在图形中用不同的颜色显示来直观地观察变量之间的关系。也可以使用箱线图来查看连续变量与离散变量之间的关系。在图形中，每个箱线图代表一个离散变量的取值。

例如，对于连续变量高峰时期通话时长与流失之间的关系，使用箱线图以体现两个变量之间的关系。

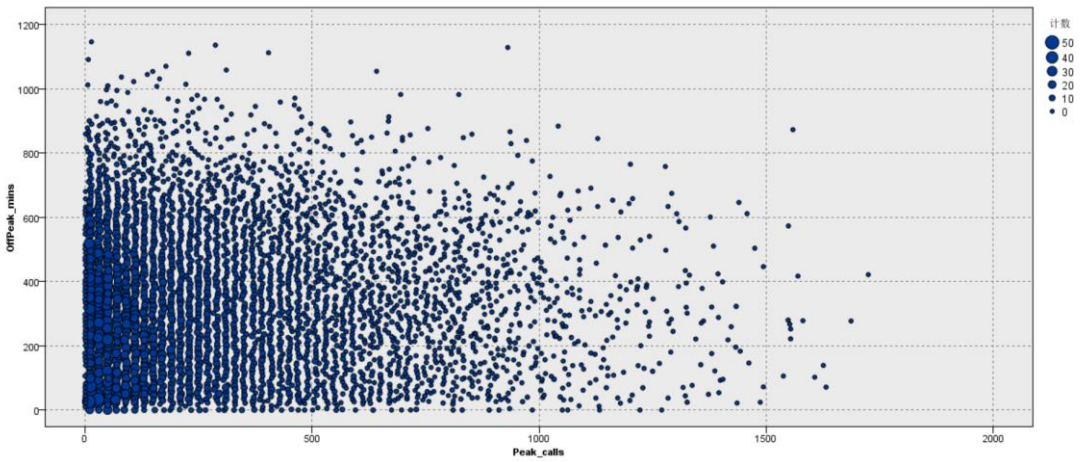


如果希望得到离散变量与连续变量之间的量化关系，则可以使用统计分析中的方差分析方法，从下图中可以看出，从统计意义上讲，在 0.05 显著性水平下。流失客户与不流失客户的高峰时期通话时长有着显著差异。

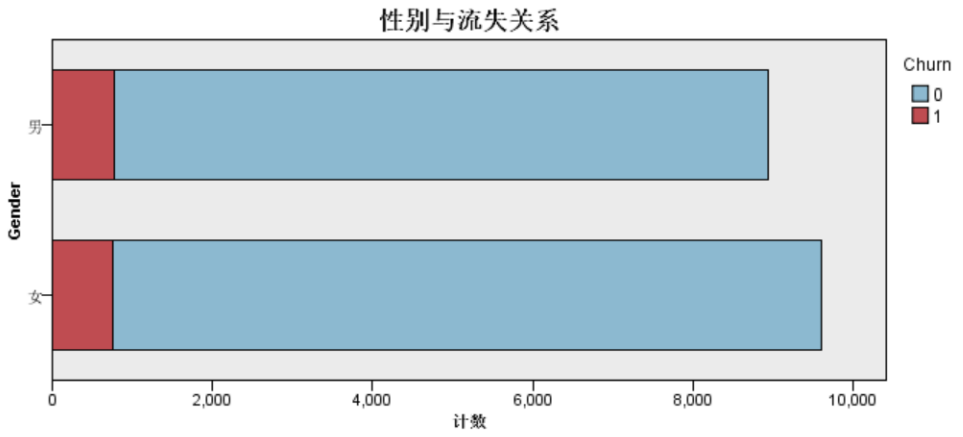
字段	0*	1*	F 检验	df	重要性
Peak_mins	715.844 504.228 3.862 17048	675.935 521.364 13.453 1502	8.599	1, 18548	0.997 重要

3) 连续变量与连续变量

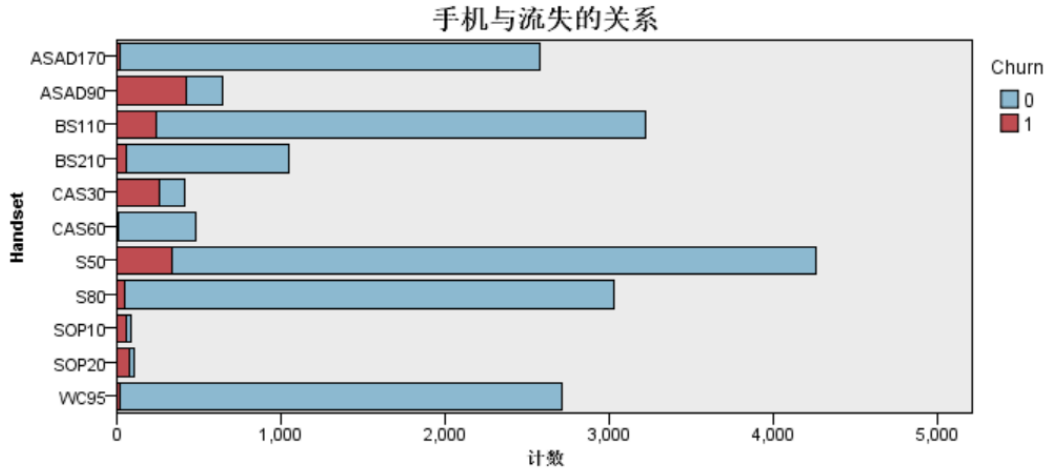
对于连续变量与连续变量之间的关系，可以使用散点图进行直观展示。例如，对于高峰时期通话数和高峰时期电话时长的关系，可以得到下图的结果：



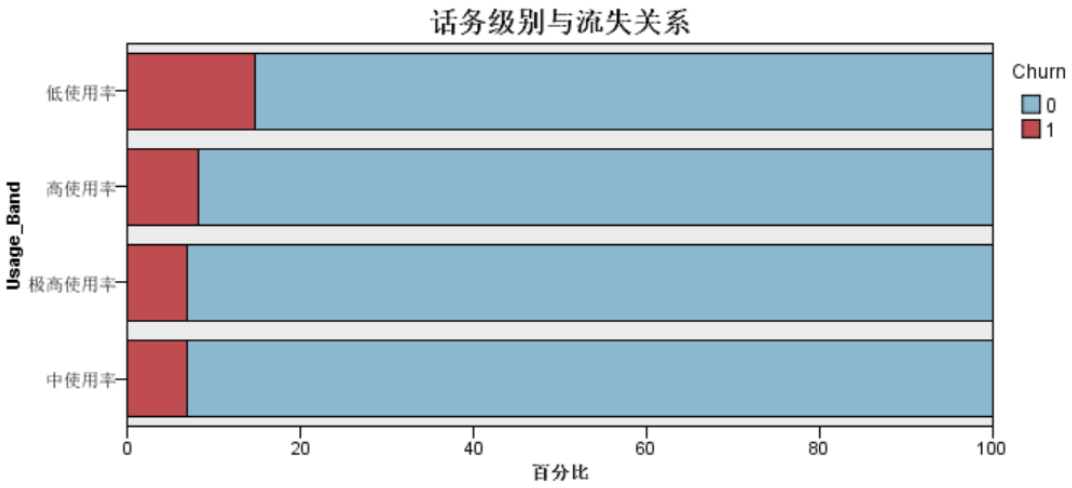
因此，在本案中，我们使用如下的分析内容：



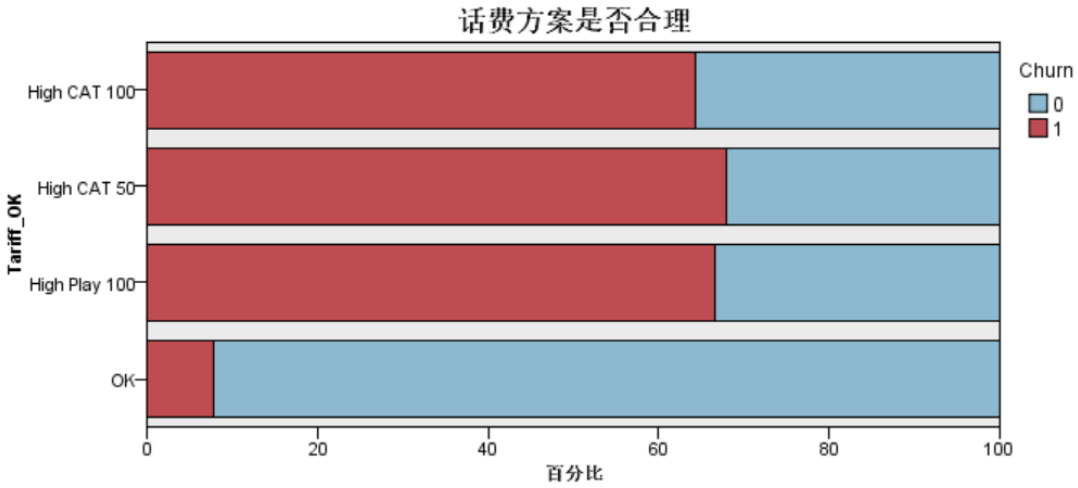
从上图可以看出，性别与流失的关系似乎不是很密切，男性和女性流失比例相差不大。



手机品牌与流失相关性很大，其中 ASAD90、CAS30、SOP10 及 SOP20 的流失比 例尤其高，猜测这些手机品牌可能使用体验(例如，信号强度、使用方便性) 较差，或许这是造成客户体验下降从而流失的根本原因，当然这只是根据数据得到的结论，实际情况如何，还需要和业务人员充分讨论，如果证实了猜测，那么或许对这些客户推荐(或赠送)其他手机品牌将是一种非常有效的挽留手段。

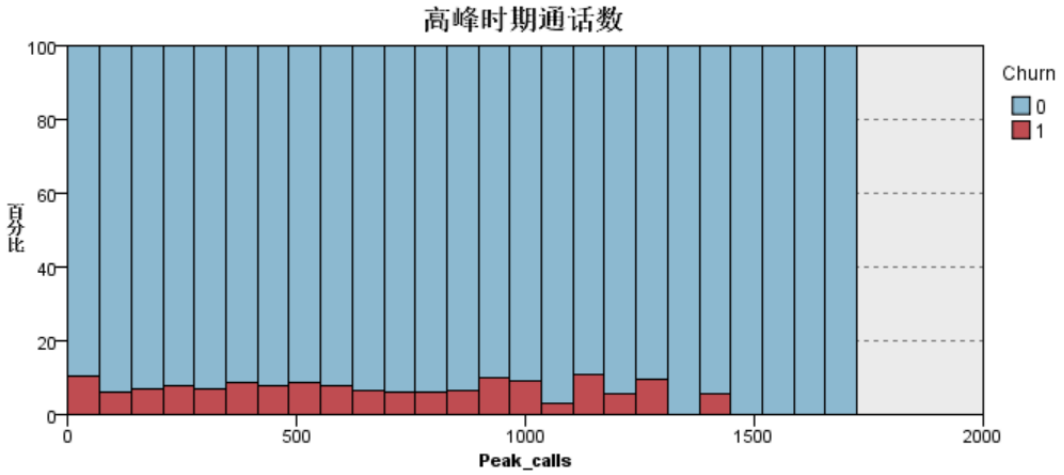


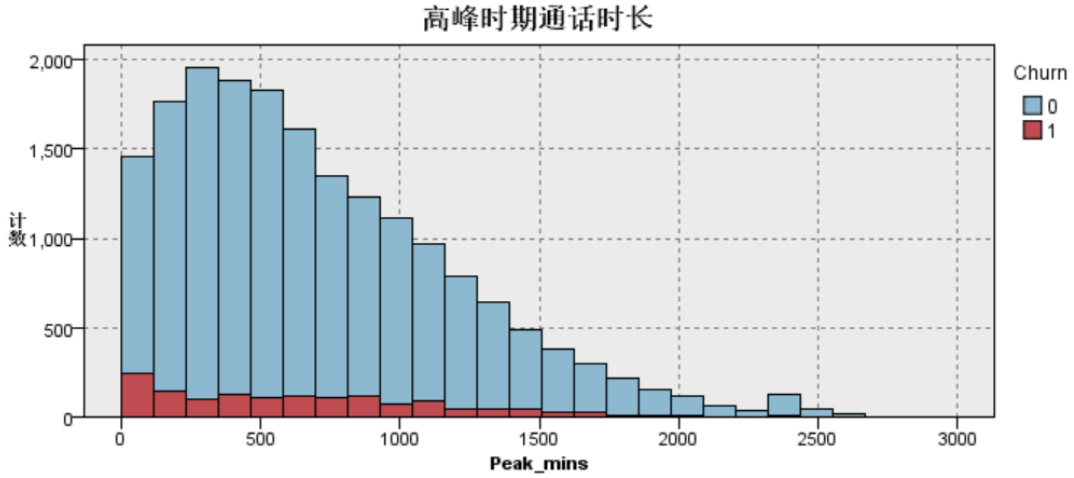
话务量级别与流失之间有一定的关系，低使用率客户流失比率要高一些，这和我们的业务经验一致。

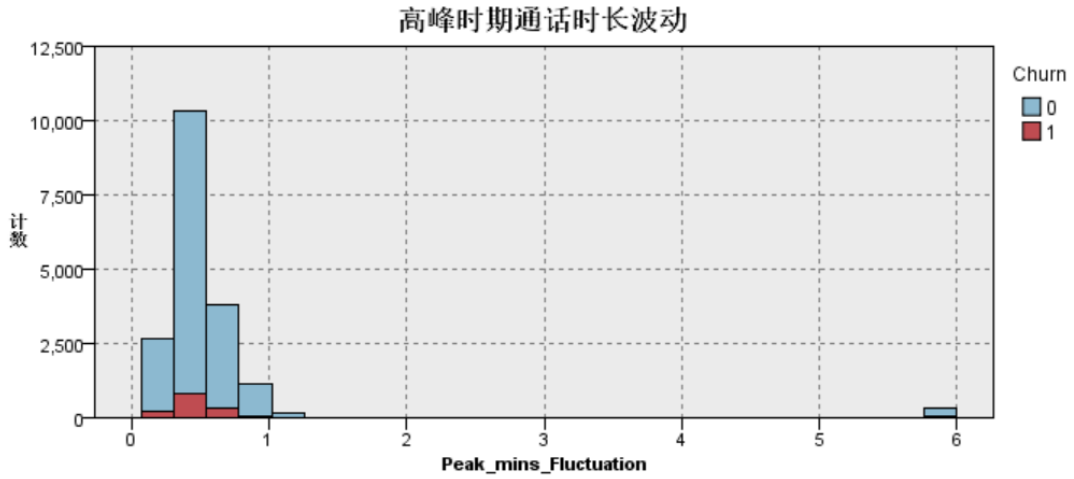


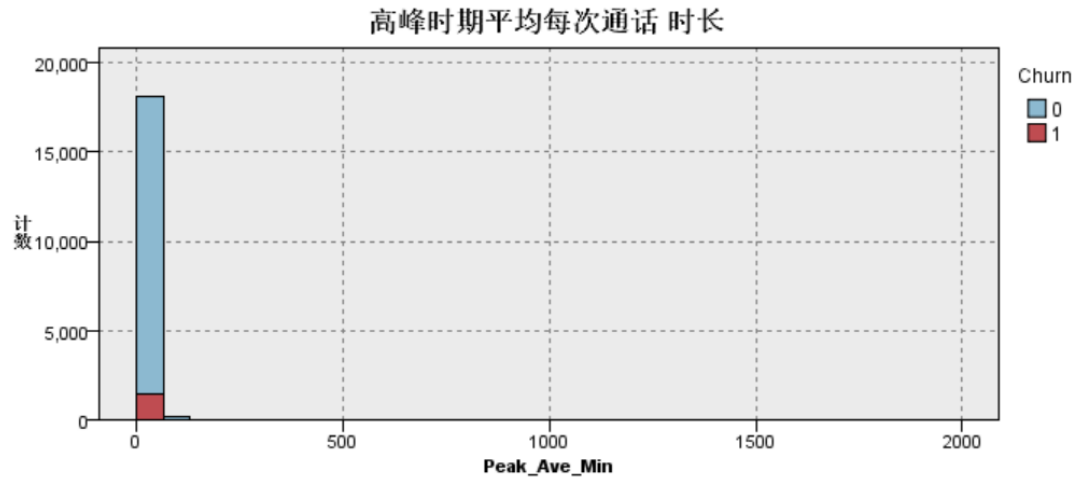
话费合理性与流失之间关系密切。在 3 种话费方案不合理(HighCAT50、HighCAT100 及 High Play100)的情形下客户流失比率远高于话费合理情形的流失比率，这提示我们的客户是非常聪明的。尽管在话费不合理情况下，可以短期内获得超额利润，但是难以长久， 可以建议业务部门关谊这一点，向客户推荐更加适合的话费方案。

对高峰时期通话行为相关的连续变量与流失之间的关系的探索性分析，得到：









流失似乎和高峰时期通话行为的关系并不是特别密切，但大致可以看出高峰时期通话时间较少、高峰时期通话时长取值很低或者很高、高峰时期通话时长被动大、高峰时期平均每次通话时长较长的客户似乎流失倾向更大一些。至于流失与各连续变量关系更

细致的分析， 我们将通过后面的建模过程来完成。