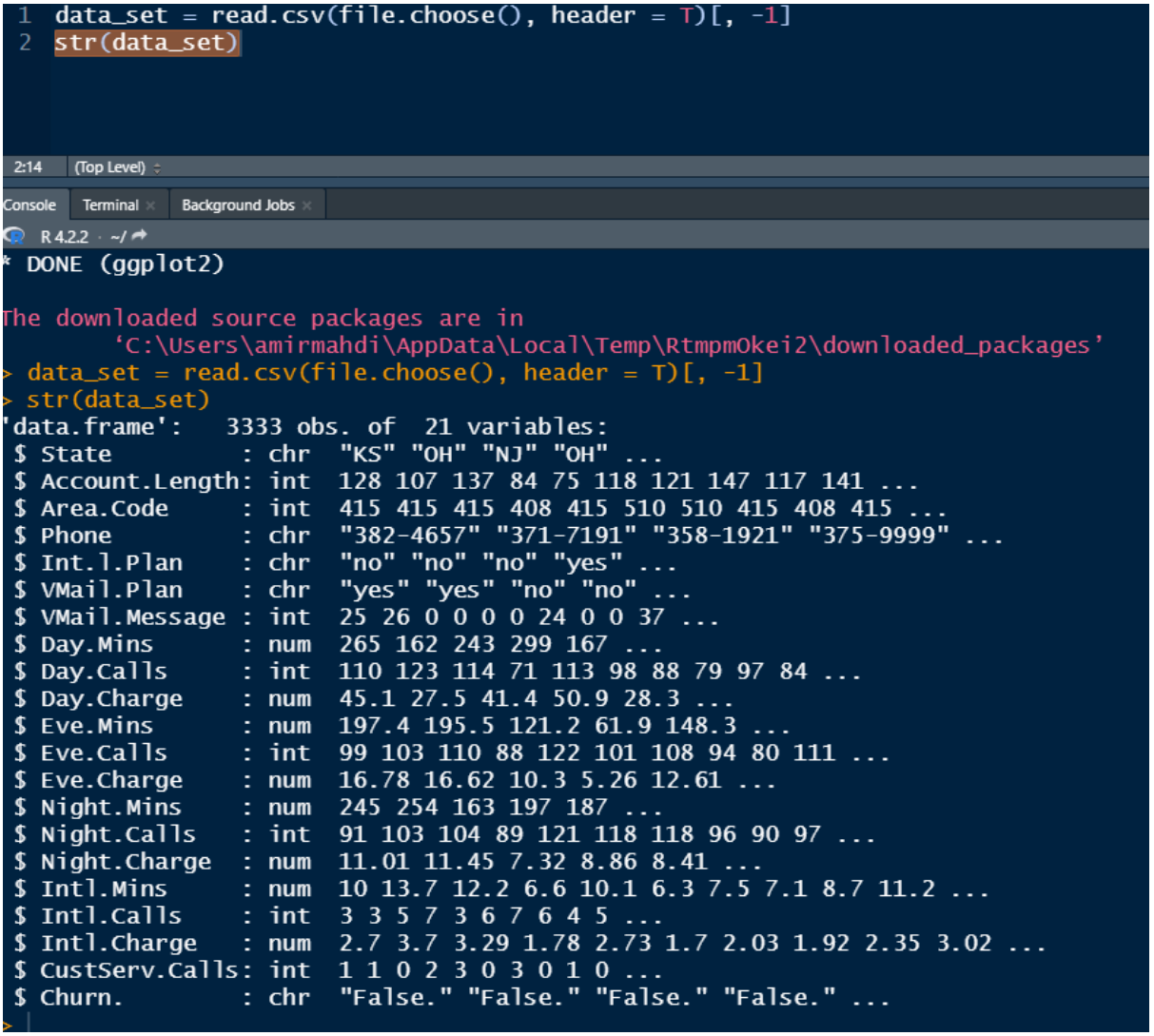
(22

Categorical: State, Int.l.Plan, Vmail.Plan, Churn.(target variable)

Numerical: Account.Length, Area.Code (more likely to be categorical), Vmail.Message, CustServ.Calls

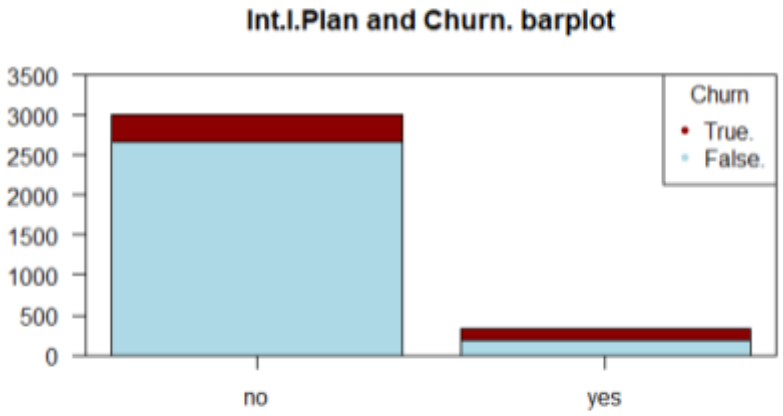
Day.Mins	Eve.Mins	Night.Mins	Intl.Mins
Day.Calls	Eve.Calls	Night.Calls	Intl.Calls
Day.Charge	Eve.Charge	Night.Charge	Intl.Charge

(دقت شود که Phone در اینجا بصورت ID رفتار میکند و بهتر است متغیر در نظر گرفته نشود و اینکه Area.code نیز اگر چه مقادیر عددی گرفته ولی هر عدد نمایانگر یک area است (رسته ای))



(25

A) Churn and Int.l.Plan



Non-normalize barchart

data_set$Churn.	data_set$Int.l.Plan no	yes	Row Total
False.	2664	186	2850
	0.885	0.576	
True.	346	137	483
	0.115	0.424	
Column Total	3010	323	3333
	0.903	0.097	

Statistics for All Table Factors

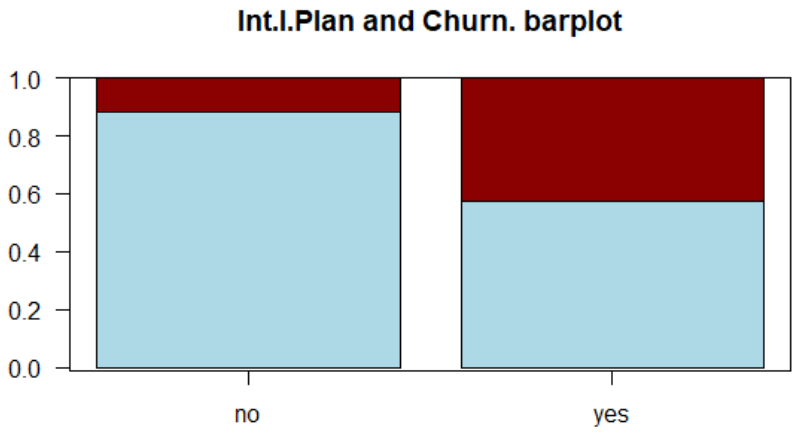
Pearson's Chi-squared test

Chi^2 = 225.0541 d.f. = 1 p = 7.14514e-51

Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 222.5658 d.f. = 1 p = 2.493108e-50

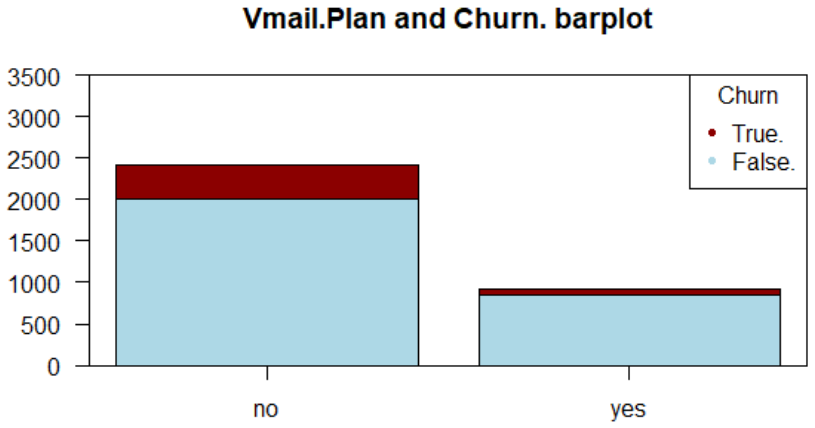
Contingency table



Normalized barchart

با بررسی barplot و contingency table میتوان نتیجه گرفت که درصد زیادی از مشتریانی که Int.l.Plan دارند، churn کردند (حدود 42%، چهار برابر کسانیکه این سرویس را ندارند). پس گویا این سرویس برای مشتریان قابل قبول نبوده. البته باید توجه کرد که فراوانی مشتریانی که Int.l.Plan دارند نسبت به آنان که ندارند نیز بسیار کم است (323 / 3333)
ضمنا p-value مقدار کمی دارد و فرض صفر (استقلال) رد میشود

B) Churn and Vmail.Plan



Non-normalized barchart

data_set$Churn.	data_set$VMail.Plan no	yes	Row Total
False.	2008	842	2850
	0.833	0.913	
True.	403	80	483
	0.167	0.087	
Column Total	2411	922	3333
	0.723	0.277	

Statistics for All Table Factors

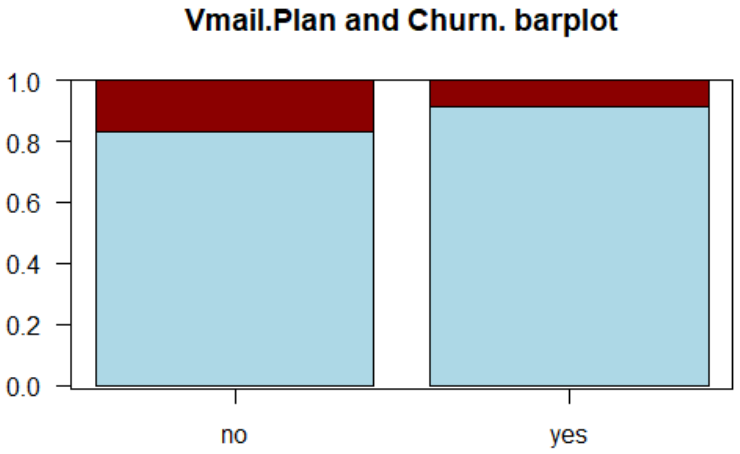
Pearson's Chi-squared test

Chi^2 = 34.77733 d.f. = 1 p = 3.696527e-09

Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 34.13166 d.f. = 1 p = 5.15064e-09

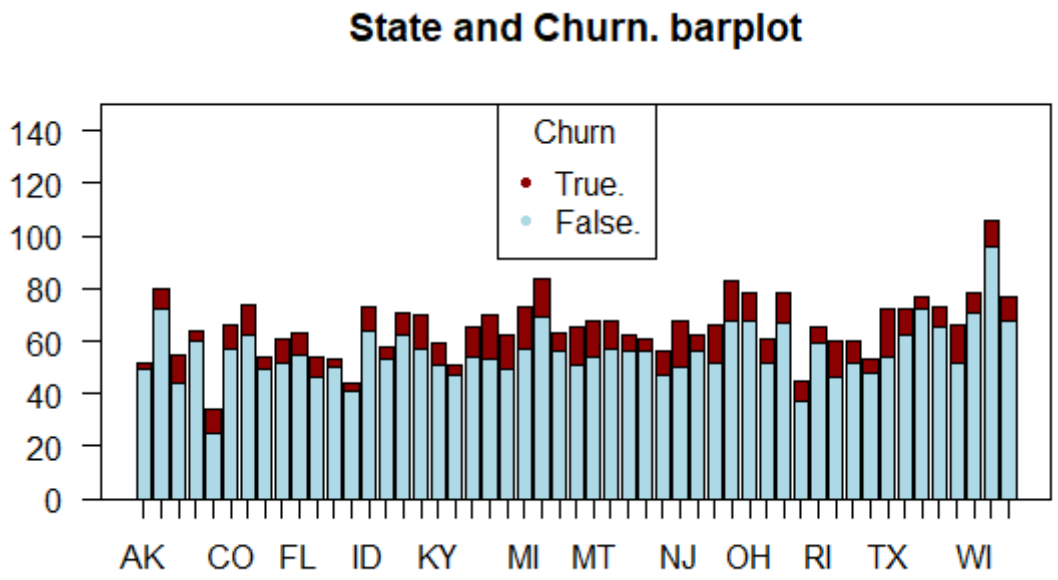
Contingency table



Normalized barchart

با بررسی barplot و contigency table میتوان نتیجه گرفت که سرویس Vmail.Plan برای مشتریان سرویس مناسبی بوده. زیرا درصد churn کنندگان در میان مشتریانی که Vmail.Plan را دارند حدود 8.7% و برای آنهایی که این سرویس را ندارند 17% می باشد.
ضمنا p-value مقدار کمی دارد و فرض صفر (استقلال) رد میشود.

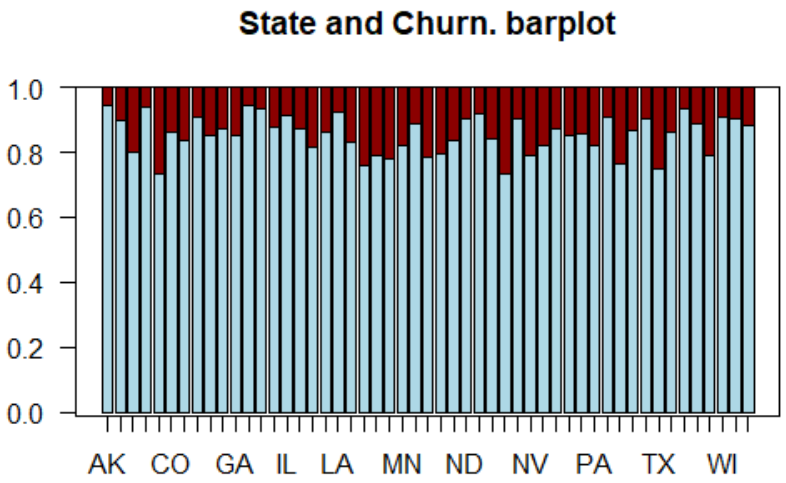
C) Churn and State



Non-normalized barchart

data_set$State	data_set$Churn. False.	True.	Row Total
AK	49 0.017	3 0.006	52
AL	72 0.025	8 0.017	80
AR	44 0.015	11 0.023	55
AZ	60 0.021	4 0.008	64
CA	25 0.009	9 0.019	34
CO	57 0.020	9 0.019	66
CT	62 0.022	12 0.025	74
DC	49 0.017	5 0.010	54
DE	52	9	61

Contingency table

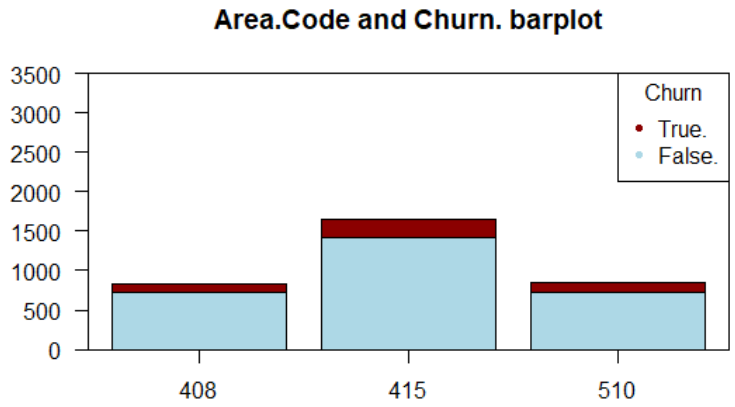


Normalized barchart

با بررسی barplot و contingency table نمیتوان با قطعیت در مورد استقلال یا عدم استقلال churn و State صحبت کرد. (p-value = 0.09).
تنها میتوان ایالت هایی که نسبت churn کنندگان در آنها به نسبت کل مشتری ها **در همان ایالت** بالا هست را یافت. (row percentage)

State	Percentage of Churn	Number of customers
CA	26%	34
MD	24%	70
NJ	26%	68
NV	21%	66
SC	23%	60
TX	25%	72
WA	21%	66

D) Churn and Area.Code



Non-Normalized barchart

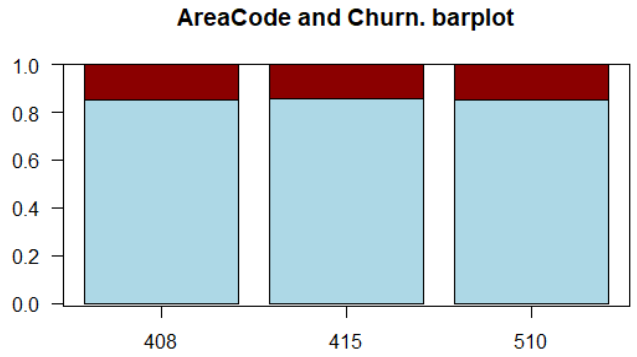
data_set$Area.Code				
data_set$Churn.	408	415	510	Row Total
False.	716	1419	715	2850
	0.854	0.857	0.851	
True.	122	236	125	483
	0.146	0.143	0.149	
Column Total	838	1655	840	3333
	0.251	0.497	0.252	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 0.1775407 d.f. = 2 p = 0.9150557

Contingency table



با بررسی barplot های بالا میبینیم که درصد Chrun در هر سه AreaCode حدود 15% است و بنظر می رسد این دو متغیر از هم مستقل اند (البته تعداد مشتریان در AreaCode = 415 حدودا 2 برابر بقیه AreaCode هاست) ضمنا p-value مقدار زیادی دارد(90%) و فرض صفر (استقلال) قبول میشود

(b

متغیر های Vmail.Plan و Int.l.Plan در هر الگوریتم data mining که در آینده قرار است بکار گرفته شود تاثیر بسزایی خواهند گذاشت

(26

ارتباط متغیر Churn با دیگر متغیر های رسته ای

	data_set$Int.l.Plan		
data_set$Churn.	no	yes	Row Total
False.	2664	186	2850
	0.885	0.576	
True.	346	137	483
	0.115	0.424	
Column Total	3010	323	3333
	0.903	0.097	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 225.0541 d.f. = 1 p = 7.14514e-51

Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 222.5658 d.f. = 1 p = 2.493108e-50

	data_set$VMail.Plan		
data_set$Churn.	no	yes	Row Total
False.	2008	842	2850
	0.833	0.913	
True.	403	80	483
	0.167	0.087	
Column Total	2411	922	3333
	0.723	0.277	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 34.77733 d.f. = 1 p = 3.696527e-09

Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 34.13166 d.f. = 1 p = 5.15064e-09

data_set$Churn.	data_set$Area.Code 408	415	510	Row Total
False.	716	1419	715	2850
	0.854	0.857	0.851	
True.	122	236	125	483
	0.146	0.143	0.149	
Column Total	838	1655	840	3333
	0.251	0.497	0.252	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 0.1775407 d.f. = 2 p = 0.9150557

data_set$State	data_set$Churn. False.	True.	Row Total
AK	49	3	52
	0.017	0.006	
AL	72	8	80
	0.025	0.017	
AR	44	11	55
	0.015	0.023	
AZ	60	4	64
	0.021	0.008	
CA	25	9	34
	0.009	0.019	
CO	57	9	66
	0.020	0.019	
CT	62	12	74
	0.022	0.025	
DC	49	5	54
	0.017	0.010	
DE	52	9	61

با بررسی درصد ها و همچنین فراوانی های crosstabulation های بالا میتوان دریافت که تفاوت درصد churn میان مشتریانی که int.l.plan دارند و ندارند بالاست (پس ازchurn مستقل نیست و pvalue نیز گویای این مطلب است). این تفاوت درصد در میان مشتریانی که vmail.plan دارند و ندارند نیز مشهود است. در مورد متغیر state بدلیل فراوانی حالت ها و زیاد نبودن p-value نمیتوان با اطمینان بالایی از مستقل بودن یا نبودن از chrun صحبت کرد و تنها میتوان گفت برخی ایالت ها درصد و فراوانی churn بالایی دارند و باید به آنها توجه کرد. متغیر Area.Code باتوجه به درصد ها و pvalue از churn با اطمینان بالایی مستقل است.

ارتباط متغیر Vmail.Plan با دیگر متغیر های رسته ای

data_set$VMail.Plan	data_set$Int.l.Plan: no	yes	Row Total
no	2180	231	2411
	0.724	0.715	
yes	830	92	922
	0.276	0.285	
Column Total	3010	323	3333
	0.903	0.097	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 0.1202429 d.f. = 1 p = 0.7287712

Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 0.07913866 d.f. = 1 p = 0.7784681

data_set$VMail.Plan	data_set$Area.Code: 408	415	510	Row Total
no	618	1184	609	2411
	0.737	0.715	0.725	
yes	220	471	231	922
	0.263	0.285	0.275	
Column Total	838	1655	840	3333
	0.251	0.497	0.252	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 1.368076 d.f. = 2 p = 0.5045755

data_set$State	data_set$VMail.Plan: no	yes	Row Total
AK	36	16	52
	0.015	0.017	
AL	59	21	80
	0.024	0.023	
AR	41	14	55
	0.017	0.015	
AZ	45	19	64
	0.019	0.021	
CA	23	11	34
	0.010	0.012	
CO	47	19	66
	0.019	0.021	
CT	53	21	74
	0.022	0.023	
DC	36	18	54
	0.015	0.020	
DE	46	15	61
	0.019	0.016	
FL	43	20	63

از cross-tabulation های بالا میتوان دریافت که با اطمینان بالا میتوان گفت متغیر Vmail-plan با متغیر های Int.l.Plan و Area.Code مستقل است (با p-value های 0.7 و 0.5) در مورد رابطه Vmail.plan و State هم علی رغم زیاد بودن حالت های state با نگاه کردن به جدول و محاسبه p-value از آزمون chi-square میتوان دریافت که این دو متغیر مستقل اند (0.98).
تنها میتوان گفت تعداد مشتریانی که vmail-plan و int.l.plan را با هم ندارند بسیار زیاد (65%)است.

ارتباط متغیر Int.l.Plan و دیگر متغیرهای رسته ای

data_set$Int.l.Plan	data_set$Area.Code: 408	415	510	Row Total
no	767	1505	738	3010
	0.915	0.909	0.879	
yes	71	150	102	323
	0.085	0.091	0.121	
Column Total	838	1655	840	3333
	0.251	0.497	0.252	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 7.936216 d.f. = 2 p = 0.01890917

data_set$State	data_set$Int.l.Plan no	yes	Row Total
AK	48	4	52
	0.016	0.012	
AL	72	8	80
	0.024	0.025	
AR	47	8	55
	0.016	0.025	
AZ	61	3	64
	0.020	0.009	
CA	30	4	34
	0.010	0.012	
CO	62	4	66
	0.021	0.012	
CT	66	8	74
	0.022	0.025	
DC	49	5	54
	0.016	0.015	
DE	51	10	61
	0.017	0.031	

حدود 45% از کل مشتریان، دارای Area.Code 405 و فاقد Int.l.Plan اند

ارتباط متغیر Area.Code با دیگر متغیر های رسته ای

data_set$State	data_set$Area.Code 408	415	510	Row Total
AK	14	24	14	52
	0.017	0.015	0.017	
AL	25	40	15	80
	0.030	0.024	0.018	
AR	13	27	15	55
	0.016	0.016	0.018	
AZ	15	36	13	64
	0.018	0.022	0.015	
CA	7	17	10	34
	0.008	0.010	0.012	
CO	25	29	12	66
	0.030	0.018	0.014	
CT	22	39	13	74
	0.026	0.024	0.015	
DC	14	27	13	54
	0.017	0.016	0.015	
DE	13	31	17	61
	0.016	0.019	0.020	
FL	12	31	20	63
	0.014	0.019	0.024	
GA	15	21	18	54
	0.018	0.013	0.021	
HI	15	30	8	53
	0.018	0.018	0.010	

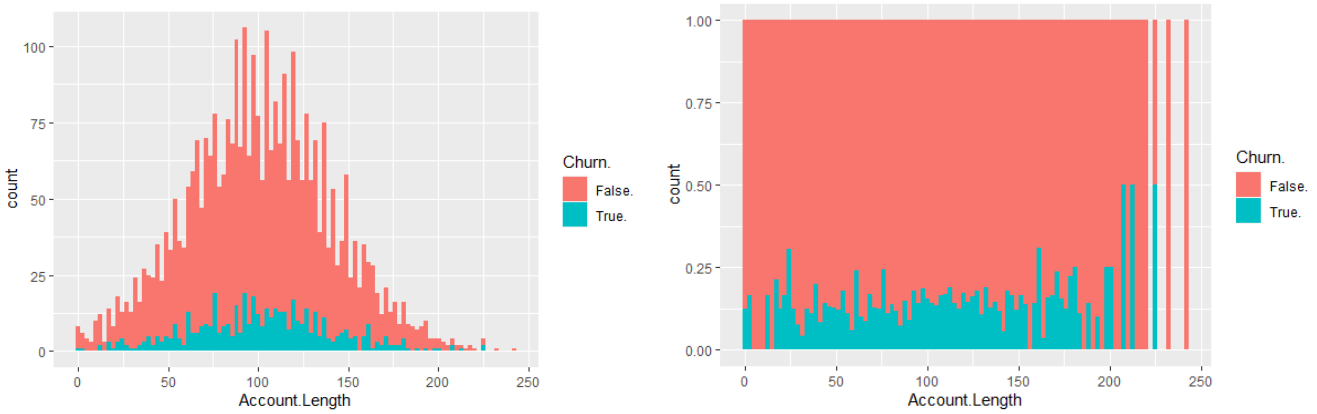
با بررسی درصد ها در هر سطر و p-value = 0.6میتوان گفت این دو متغیر مستقلند.

(28

مشتریانی که Churn کرده اند در AreaCode های مختلف بطور تقریبا یکسانی توزیع شده اند.(p-value = 0.91)
همچنین در هر State مشتریان با AreaCode های مختلف نیز بطور تقریبا یکسان توزیع شده اند. (= p_value 0.6) همچنین متغیر Vmail_plan با متغیر های State و Int.l.Plan و Area.Code با مقادیر = p-value 0.5 ,0.77 ,0.98 نیز حاکی از این است که این متغیر ها تاثیری بر مشتریان در گرفتن یا نگرفتن Vmail.Plan ندارند (مستقل از Vmail.Plan اند)
روابط بالا anomalous اند و باید قبل از مرحله ورودی دادن به الگوریتم با یک فرد expert در مورد این متغیرها صحبت کرد یا صحت data set را سوال کرد.

(30

بررسی متغیر Account.length

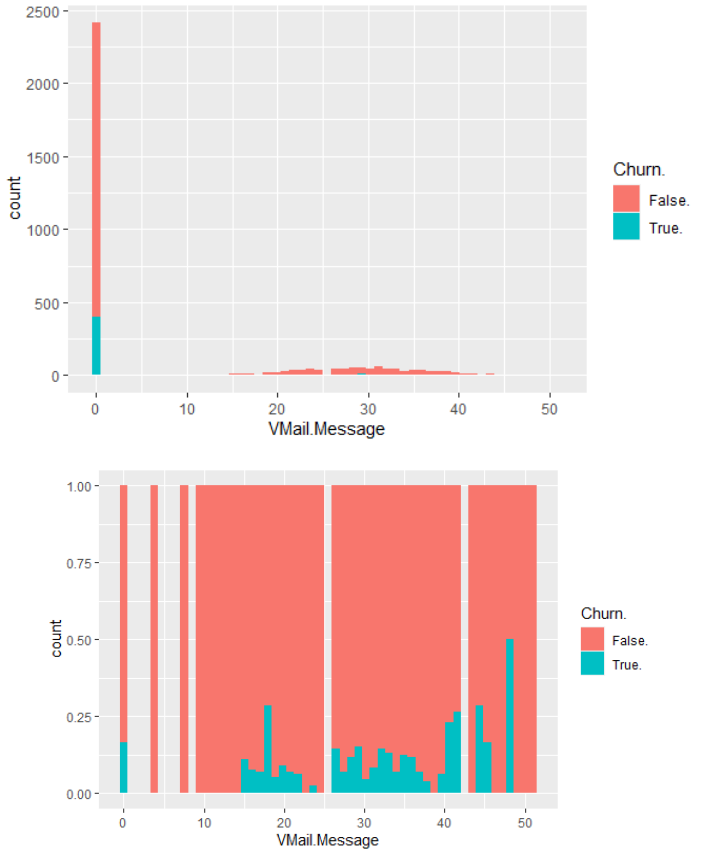



```
19  churn.false = subset(data_set, data_set$Churn. == "False.")
20  churn.true = subset(data_set, data_set$Churn. == "True.")
21
22  t.test(churn.false$Account.Length, churn.true$Account.Length)
23
24
25
26
23:1  (Top Level) :

Console  Terminal  Background Jobs
R 4.2.2 · ~/
> churn.false = subset(data_set, data_set$Churn. == "False.")
> churn.true = subset(data_set, data_set$Churn. == "True.")
>
> t.test(churn.false$Account.Length, churn.true$Account.Length)

        Welch Two Sample t-test

data:  churn.false$Account.Length and churn.true$Account.Length
t = -0.96189, df = 659.91, p-value = 0.3365
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.690123  1.948299
sample estimates:
mean of x mean of y
 100.7937  102.6646
```

T-test for Churn and Account.Length relation

مقدار p-value نشان میدهد ه میتوان با اطمینان خوبی گفت متغیر Account.Length از Churn مستقل است

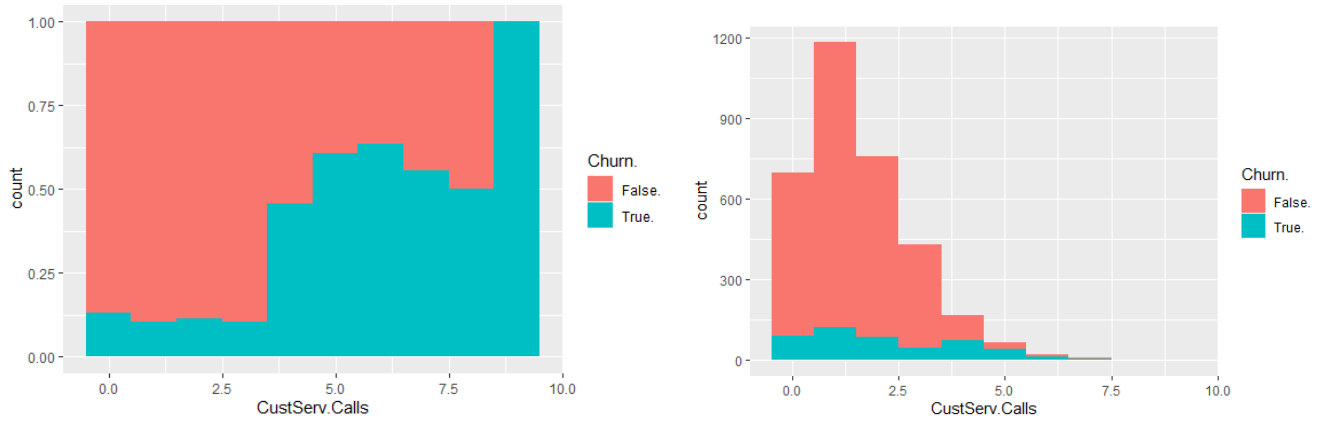
بررسی متغیر Vmail.Message



pvalue = 8.765e-09

تعداد بسیار زیادی از مشتریان این سرویس را ندارند که درصد churn نسبتا پایینی هم دارند.

بررسی متغیر CustServ.Calls



p-value < 2.2e-16

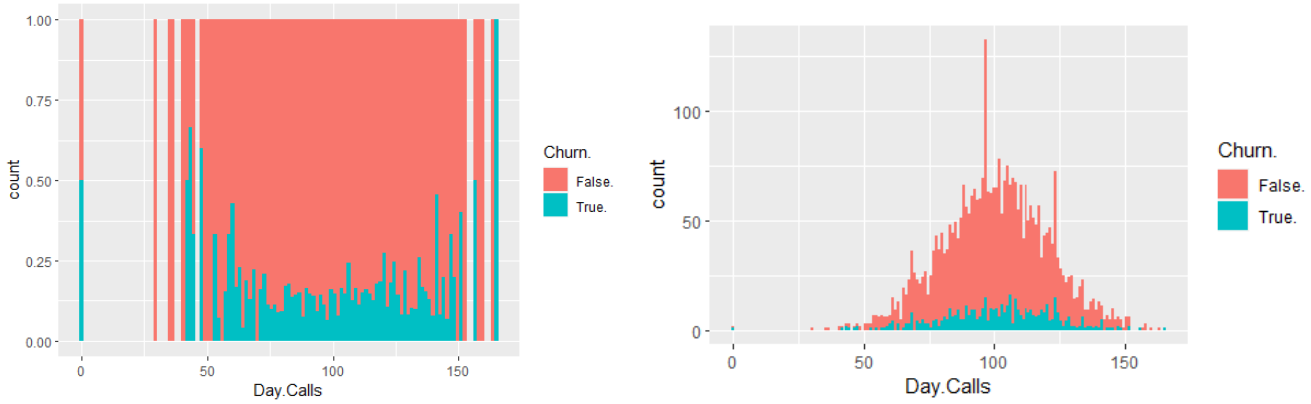
با افزایش custServ.Calls، درصد churn افزایش می یابد (البته فراوانی مشتریان با custServ.Calls بالا نیز کم است)

بررسی متغیر Day.Mins

p-value < 2.2e-16

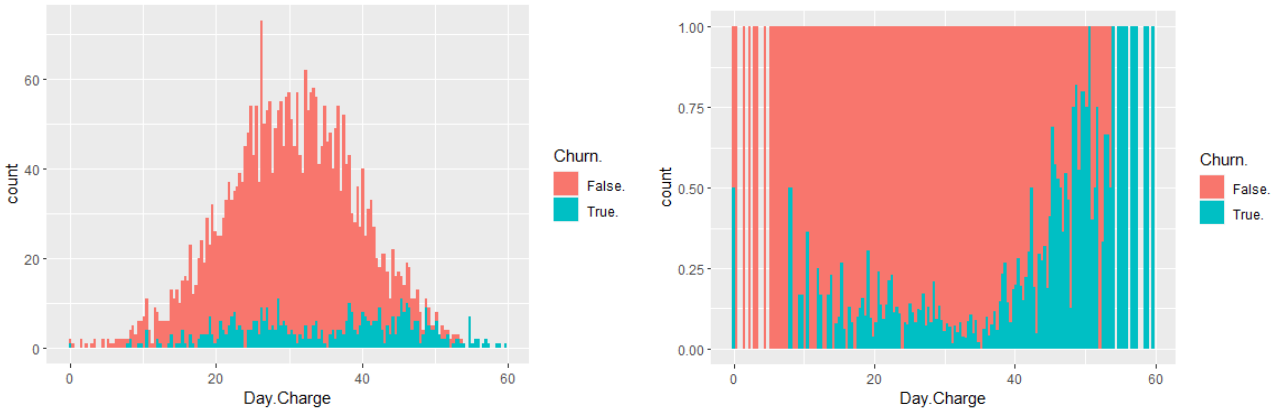
مشتریان با DayMins بیشتراز 250 احتمال churn بالایی دارند (البته اکثریت مشتریان در این بازه نیستند)

بررسی متغیر Day.Calls



p-value: 0.3165

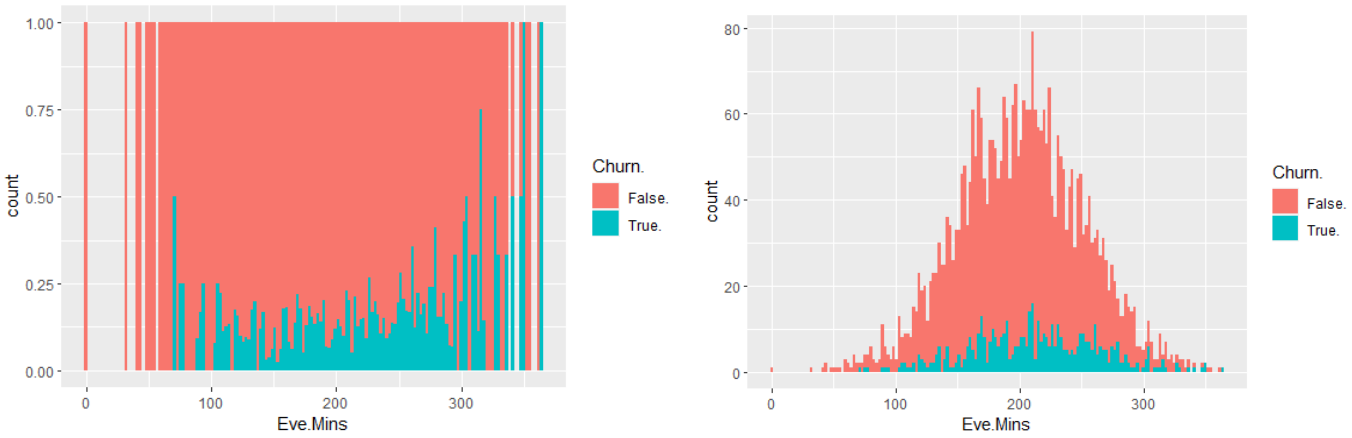
بررسی متغیر Day.Charge



P-value < 2.2e-16

این متغیر با متغیر Day.Mins ارتباط مستقیم دارد (cor(Day.Charge, Day.Mins)v = 1 حاکی این مطلب است). مشتریان بالای 45 دلار Day.Charge، فراوانی کم و درصد churn بالایی دارند.

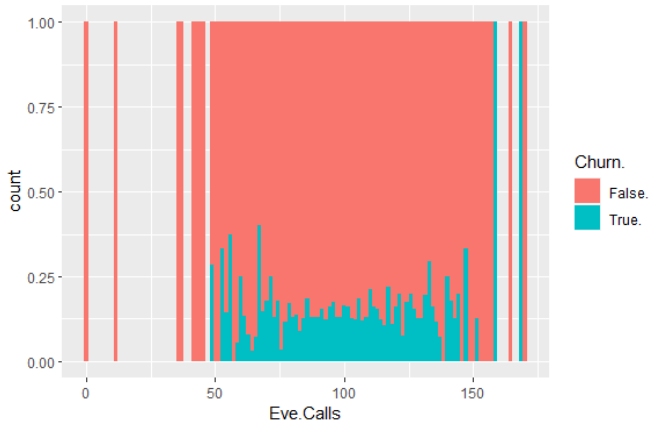
بررسی متغیر EVE.Mins

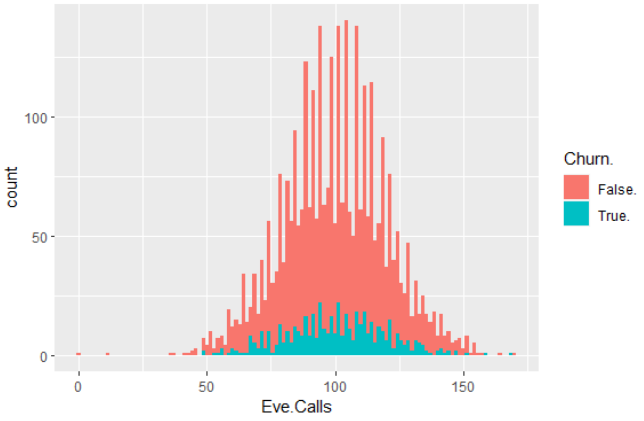


p-value: 1.839e-07

مشتریان بالای 300 Eve.Mins دارای churn rate بالایی هستند

بررسی متغیر EVE.Calls:

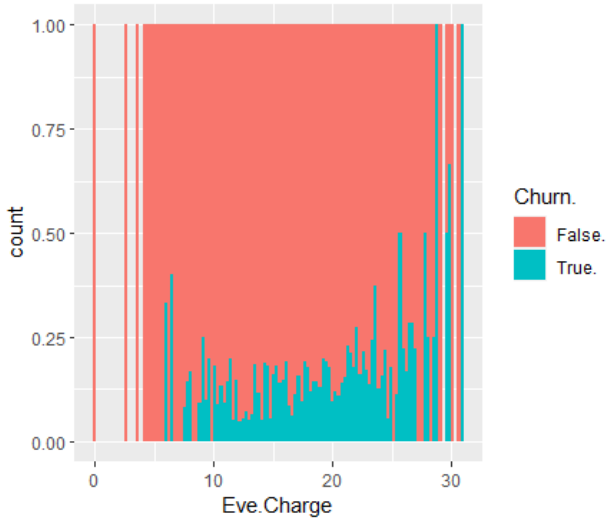


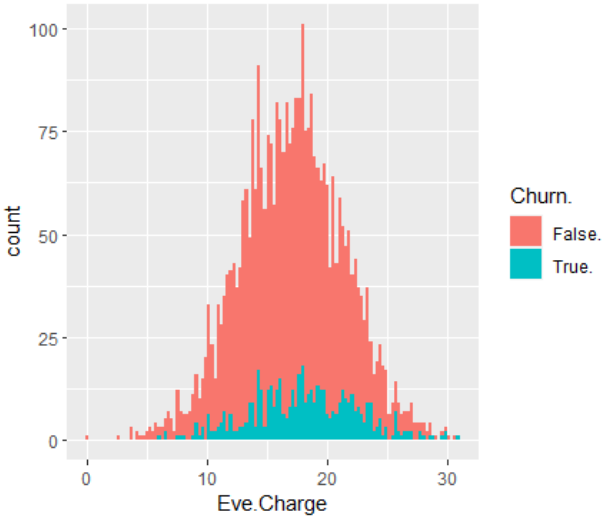


p-value = 0.5912

با اطمینان بالای میتوان گفت Eve.Calls و churn مستقل اند

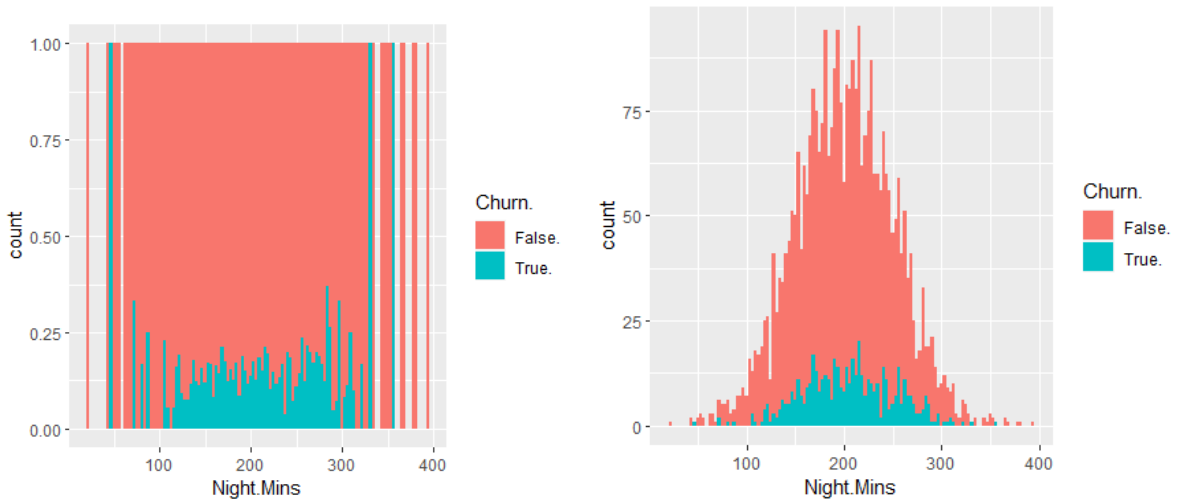
بررسی متغیر EVE.Charge





p-value: 1.843e-07

بررسی متغیر Nigth.Mins

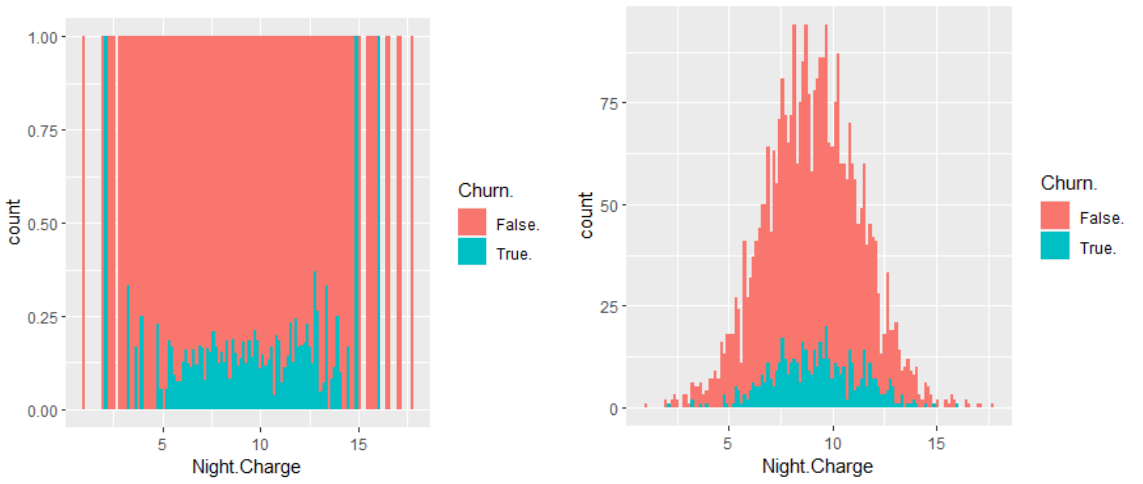


p-value: 0.03028

بررسی متغیر Night.Calls

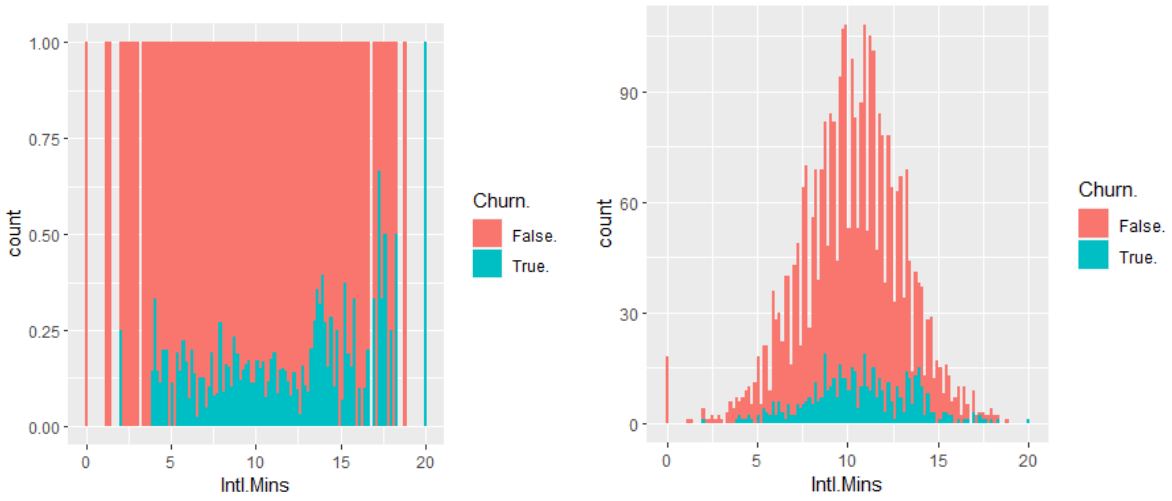
p-value = 0.7273

بررسی متغیر Night.Charge



p-value: 0.03027

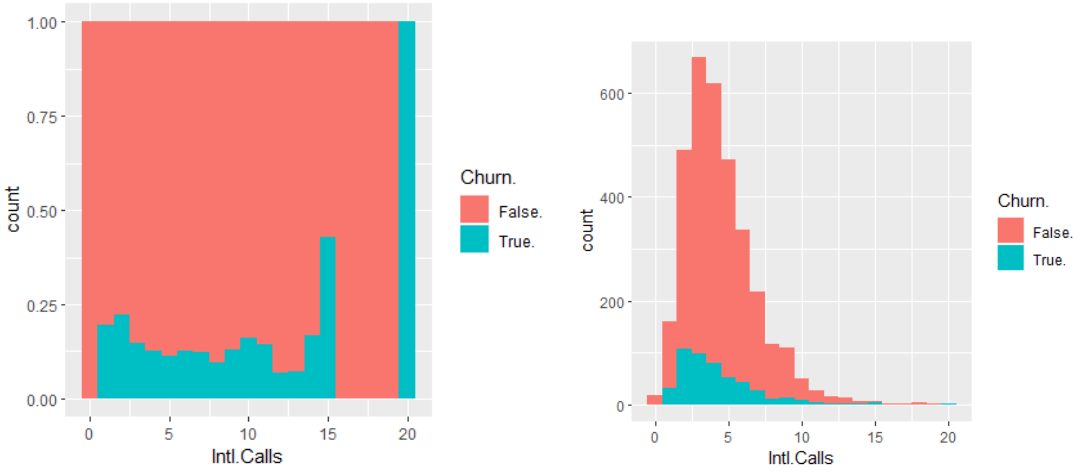
بررسی متغیر Intl.Mins:



p-value: 9.066e-05

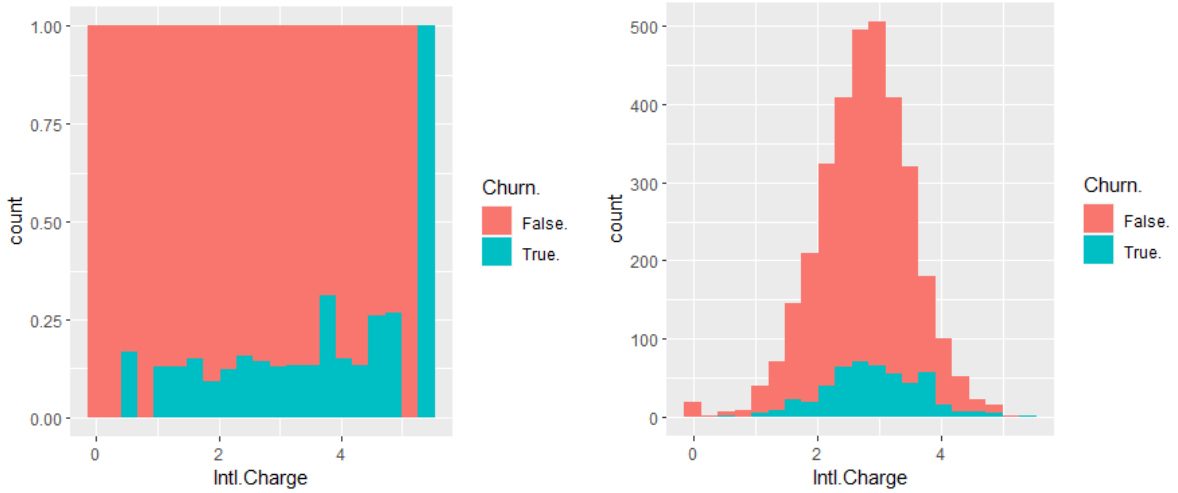
مشتریان دارای Intl.Mins بالای 15 Churn بالا و فراوانی کمی دارند

بررسی متغیر Intl.Calls:



p-value: 0.003186

بررسی متغیر Intl.Charge



p-value: 9.026e-05

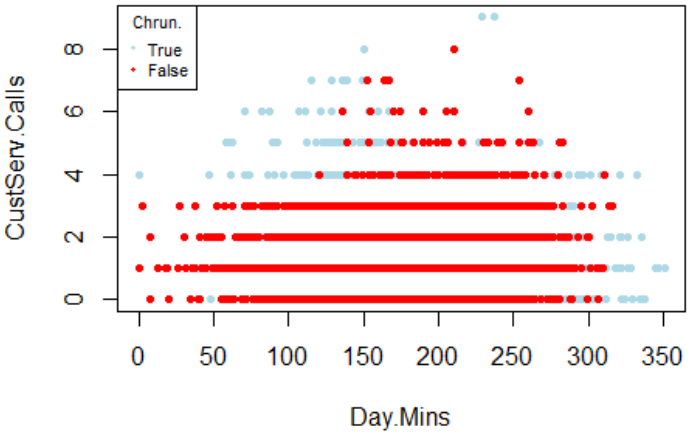
(b

با مشاهده p-value ها متوجه میشویم که احتمالا متغیرهای مربوط به Calls غیر international (یعنی Eve/Day/Night Calls) با churn مستقل اند و تاثیر بسزایی در الگوریتم ما نخواهند گذاشت (البته آنها را حذف نمیکنیم، بلکه برای مراحل بعد کاندید میکنیم).

با محاسبه cor بین متغیر های عددی فوق متوجه میشویم که بین هر جفت Mins و Charge رابطه تقریبا خطی وجود دارد (در تمامی جفت متغیرهای Day.Mins و Day.Charge یا Night.Mins و Night.Charge و ...) پس بهتر است از هر جفت، یکی از متغیرها را در الگوریتم لحاظ کنیم.

بقیه متغیر ها را باید در الگوریتم مان لحاظ کنیم

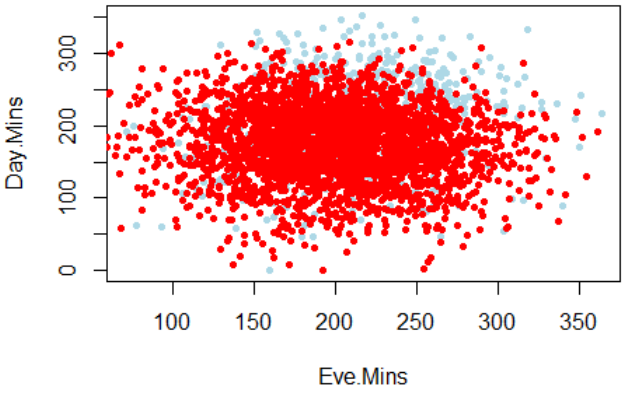
(32




```
18
19  churn.false = subset(data_set, data_set$Churn. == "False." &
20                         data_set$Day.Mins < 150 & data_set$CustServ.Calls >= 4)
21  churn.true = subset(data_set, data_set$Churn. == "True." &
22                        data_set$Day.Mins < 150 & data_set$CustServ.Calls >= 4)
23
24  t.test(churn.true$Intl.Charge, churn.false$Intl.Charge)
25
26
16:55  (Top Level)                                                    R Scri
Console  Terminal  Background Jobs
R 4.2.2 · ~/
+                         data_set$Day.Mins < 150 & data_set$CustServ.Calls >= 4)
> churn.true = subset(data_set, data_set$Churn. == "True." &
+                       data_set$Day.Mins < 150 & data_set$CustServ.Calls >= 4)
> t.test(churn.true$Intl.Charge, churn.false$Intl.Charge)

        Welch Two Sample t-test

data:  churn.true$Intl.Charge and churn.false$Intl.Charge
t = -0.31876, df = 7.652, p-value = 0.7584
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6491370  0.4925655
sample estimates:
mean of x mean of y
 2.724571  2.802857
```

از سوال 30 میتوان دریافت که مشتریان با CustServ.Calls بالای 3 churn rate بالایی دارند. اما وقتی این متغیر با Day.Mins بصورت همزمان بررسی میشوند میتوان دریافت که آن دسته از مشتریان که CustServ.Calls بیشتر از 3 دارند اما Day.Mins بالای 150 هم دارند کمتر churn میکنند.

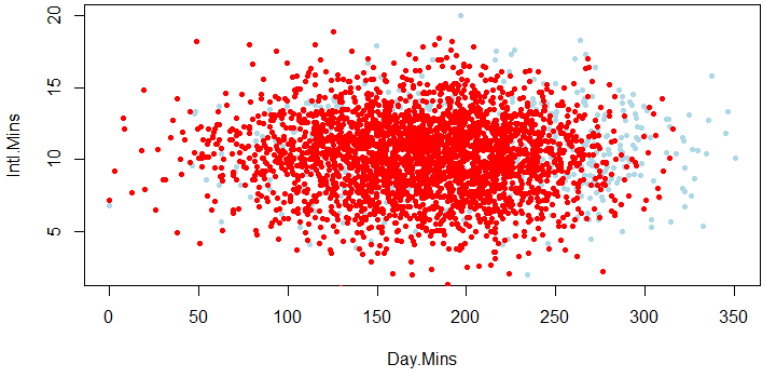
از طرفی آن دسته از مشتریانی که CustServ.Calls کمتر از 3 و Day.Mins شان از 300 بالاتر است churn rate بالایی دارند.




```
19 churn.false = subset(data_set, data_set$Churn. == "False." &
20                         data_set$Day.Mins > 250 & data_set$Eve.Mins >= 200)
21 churn.true = subset(data_set, data_set$Churn. == "True." &
22                        data_set$Day.Mins > 250 & data_set$Eve.Mins >= 200)
23
24 t.test(churn.true$Intl.Charge, churn.false$Intl.Charge)
25
26
25:1  (Top Level)

Console  Terminal  Background Jobs
R 4.2.2 · ~/

+                        data_set$Day.Mins > 250 & data_set$Eve.Mins >= 200)
> churn.true = subset(data_set, data_set$Churn. == "True." &
+                       data_set$Day.Mins > 250 & data_set$Eve.Mins >= 200)
> t.test(churn.true$Intl.Charge, churn.false$Intl.Charge)

        Welch Two Sample t-test

data:  churn.true$Intl.Charge and churn.false$Intl.Charge
t = -0.38317, df = 133.36, p-value = 0.7022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2559487  0.1728745
sample estimates:
mean of x mean of y
 2.793540  2.835077
```

از سوال 30 میتوان دریافت که مشتریان با Day.Mins بالای 250 churn بالایی دارند، اما آن دسته از مشتریان با Day.Mins بالای 250 و Eve.Mins کمتر از 200 churn بالایی ندارند. در نتیجه مشتریان با Day.Mins و Eve.Mins بالا بطور قابل توجهی churn rate بالایی دارند.

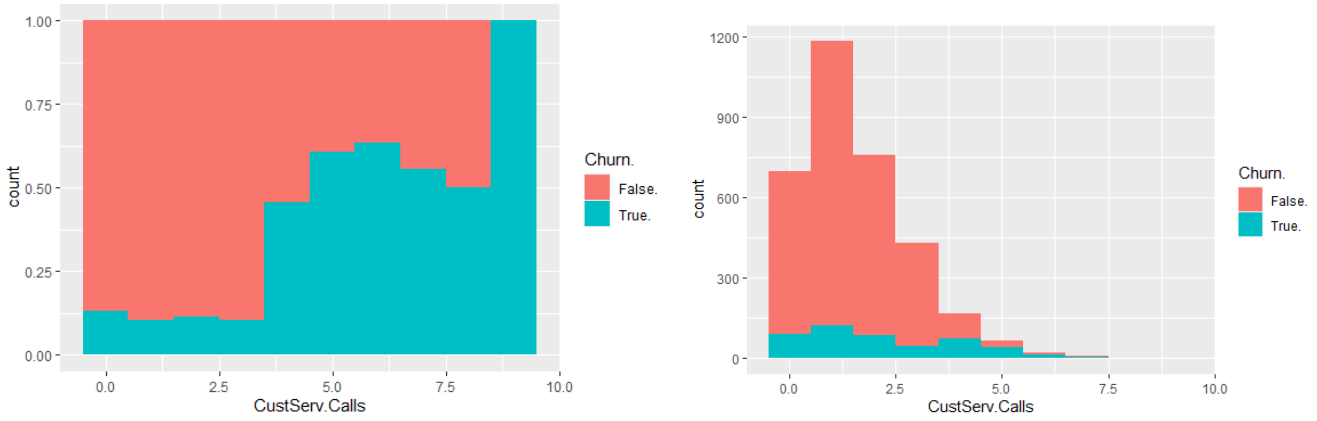
از آنجا که رابطه مستقیمی بین Eve.Mins و Eve.Charge وجود دارد، در نتیجه نتایج فوق برای نمودار Day.Mins و Eve.Charge نیز برقرار است.




```
19  churn.false = subset(data_set, data_set$Churn. == "False." &
20                         data_set$Day.Mins > 200 & data_set$Intl.Mins >= 13)
21  churn.true = subset(data_set, data_set$Churn. == "True." &
22                        data_set$Day.Mins > 200 & data_set$Intl.Mins >= 13)
23
24  t.test(churn.true$Intl.Charge, churn.false$Intl.Charge)
25
26
25:1  (Top Level)

Console  Terminal  Background Jobs
R 4.2.2 · ~/
+                         data_set$Day.Mins > 200 & data_set$Intl.Mins >= 13)
> churn.true = subset(data_set, data_set$Churn. == "True." &
+                       data_set$Day.Mins > 200 & data_set$Intl.Mins >= 13)
> t.test(churn.true$Intl.Charge, churn.false$Intl.Charge)

        Welch Two Sample t-test

data:  churn.true$Intl.Charge and churn.false$Intl.Charge
t = 1.3857, df = 87.634, p-value = 0.1694
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03280006  0.18386589
sample estimates:
mean of x mean of y
 3.896897  3.821364
```

از سوال 30 بنظر می رسید آن دسته مشتریانی که Int.Mins بالای 13 دارند با احتمال بیشتری Churn میکنند. اما با نگاه به ترکیب این متغیر با Day.Mins میتوان متوجه شد که مشتریان با Intl.Mins بالای 13 و Day.Mins بالای 200 هستند که درصد بیشتر churn کنندگان حالت univariate را تشکیل میدهند.

(33



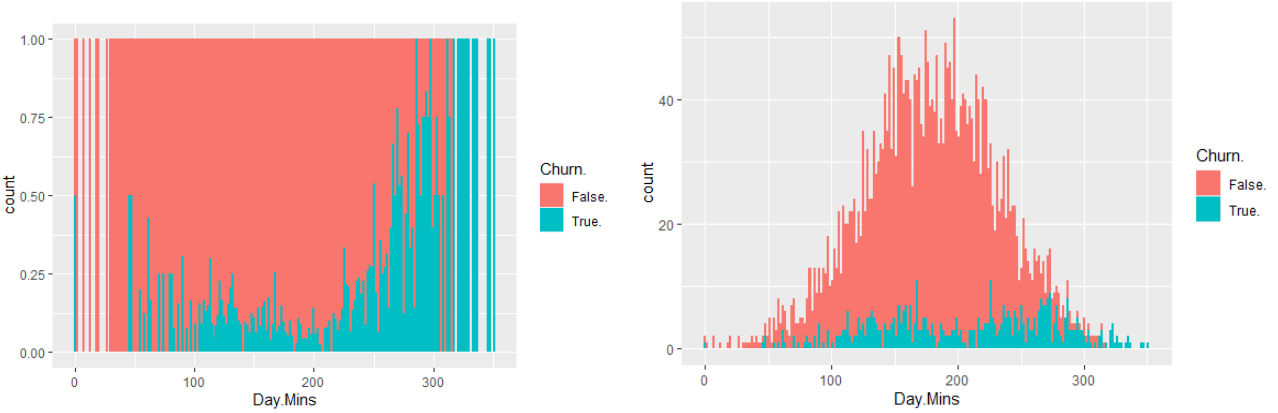
مثال کتاب: تقسیم CustServCalls به دو دسته low (کمتر از 4) و high (بیشتر از 4).

```
29 new_data_set = cbind(CustServ = data_set$CustServ.Calls,
30                      ChurnRate = data_set$Churn.)
31
32 # nrow(new_data_set)
33 # as.numeric(new_data_set[1, 1])
34
35 for(i in 1 : nrow(new_data_set)){
36   if(as.numeric(new_data_set[i, 1]) >= 4){
37     new_data_set[i, 1] = "high";
38   }
39   else{
40     new_data_set[i, 1] = "low";
41   }
42 }
43
44 library(gmodels)
45 CrossTable(y = new_data_set[, 1], x = new_data_set[, 2],
46            prop.r = FALSE, prop.t = FALSE
47            , chisq = TRUE, prop.chisq = FALSE,
48            dnn = c("Churn", "Cust.Serv.Calls"))
```

```
40:32  (Untitled) :

Console  Terminal  Background Jobs

R 4.2.2 · ~/

             | Cust.Serv.Calls
       Churn |      high |       low | Row Total |
-------------|-----------|-----------|-----------|
      False. |       129 |      2721 |      2850 |
             |     0.483 |     0.887 |           |
-------------|-----------|-----------|-----------|
       True. |       138 |       345 |       483 |
             |     0.517 |     0.113 |           |
-------------|-----------|-----------|-----------|
Column Total |       267 |      3066 |      3333 |
             |     0.080 |     0.920 |           |
-------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  324.0392     d.f. =  1     p =  1.910278e-72
```

درصد churn در این تقسیم بندی جدید حدود 4 برابر (0.517 / 0.113) می باشد و شرط استقلال قویا رد میشود که نشان میدهد تقسیم بندی خوبی است (Maximize the effect of classes)

یک مثال دیگر




```
new_data_set = cbind(CustServ = data_set$Day.Mins,
                     ChurnRate = data_set$Churn.)

# nrow(new_data_set)
# as.numeric(new_data_set[1, 1])

for(i in 1 : nrow(new_data_set)){
  if(as.numeric(new_data_set[i, 1]) >= 250){
    new_data_set[i, 1] = "high";
  }
  else{
    new_data_set[i, 1] = "low";
  }
}

library(gmodels)
CrossTable(y = new_data_set[, 1], x = new_data_set[, 2],
           prop.r = FALSE, prop.t = FALSE
           , chisq = TRUE, prop.chisq = FALSE,
           dnn = c("Churn", "Day.Mins"))
```

```
Total Observations in Table:  3333


             | Day.Mins
       Churn |      high |       low | Row Total |
-------------|-----------|-----------|-----------|
      False. |       174 |      2676 |      2850 |
             |     0.530 |     0.891 |           |
-------------|-----------|-----------|-----------|
       True. |       154 |       329 |       483 |
             |     0.470 |     0.109 |           |
-------------|-----------|-----------|-----------|
Column Total |       328 |      3005 |      3333 |
             |     0.098 |     0.902 |           |
-------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  309.3387     d.f. =  1     p =  3.042938e-69
```

تقسیم بندی Day.Mins به دو دسته low (کمتر از 250)و high (بشتر از 250)
درصد churn کردن Day.Mins = high حدود 5 برابر Day.Mins = low می باشد. ضمنا p-value نیز استقلال را رد میکند (Maximize the effect of classes)

حال تقسیم بندی را طوری انجام میدهیم که اثر کلاس ها minimum شوند (یعنی با این کلاس بندی متغیر Day.Mins و Churn مستقل میشوند)

```
new_data_set = cbind(CustServ = data_set$Day.Mins,
                     ChurnRate = data_set$Churn.)
for(i in 1 : nrow(new_data_set)){
  if(as.numeric(new_data_set[i, 1]) <= 100){
    new_data_set[i, 1] = "low";
  }
  else{
    new_data_set[i, 1] = "high";
  }
}

CrossTable(y = new_data_set[, 1], x = new_data_set[, 2],
           prop.r = FALSE, prop.t = FALSE
           , chisq = TRUE, prop.chisq = FALSE,
           dnn = c("Churn", "Day.Mins"))
```

```
Console  Terminal  Background Jobs
R 4.2.2 · ~/

             | Day.Mins
       Churn |      high |       low | Row Total |
-------------|-----------|-----------|-----------|
      False. |      2639 |       211 |      2850 |
             |     0.853 |     0.883 |           |
-------------|-----------|-----------|-----------|
       True. |       455 |        28 |       483 |
             |     0.147 |     0.117 |           |
-------------|-----------|-----------|-----------|
Column Total |      3094 |       239 |      3333 |
             |     0.928 |     0.072 |           |
-------------|-----------|-----------|-----------|

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  1.601107     d.f. =  1     p =  0.2057464
```

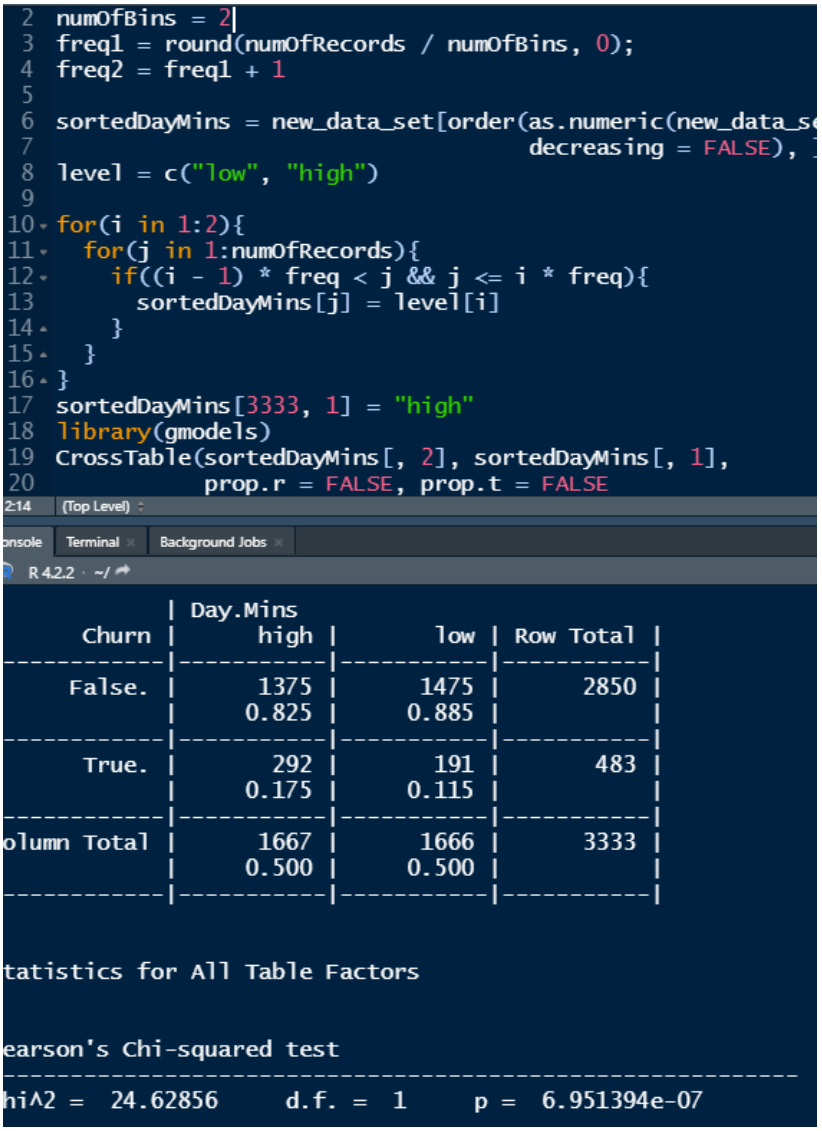
همانطور که دیده میشود، اگر مرز میان low و high را در Day.Mins مساوی 100 قرار دهیم، بخوبی نتوانسته ایم تفاوت churn rate را در تقسیم بندی مان نشان دهیم (churn rate ها در هر ستون مقدار تقریبا برابری با سطر نظیرشان دارند و p-value بزرگتر از 0.05)

(34

تقسیم بندی به دو دسته high و low بر اساس روش equal-frequency binning

دقت شود که باید data_set را قبل از binning مرتب (sort) کنیم

همانطور که دیده میشود، تفاوت قابل توجهی میان درصد churn کنندگان و آنانی که churn نمیکنند در DayMins = high / low دیده نمیشود. پس تقسیم بندی مناسبی نیست



```
numOfBins = 2
freq1 = round(numOfRecords / numOfBins, 0);
freq2 = freq1 + 1

sortedDayMins = new_data_set[order(as.numeric(new_data_s
                                   decreasing = FALSE),
level = c("low", "high")

for(i in 1:2){
  for(j in 1:numOfRecords){
    if((i - 1) * freq < j && j <= i * freq){
      sortedDayMins[j] = level[i]
    }
  }
}
sortedDayMins[3333, 1] = "high"
library(gmodels)
CrossTable(sortedDayMins[, 2], sortedDayMins[, 1],
           prop.r = FALSE, prop.t = FALSE
```

```
2:14  (Top Level)
onsole  Terminal  Background Jobs
R 4.2.2 · ~/

             | Day.Mins
       Churn |      high |       low | Row Total |
-------------|-----------|-----------|-----------|
      False. |      1375 |      1475 |      2850 |
             |     0.825 |     0.885 |           |
-------------|-----------|-----------|-----------|
       True. |       292 |       191 |       483 |
             |     0.175 |     0.115 |           |
-------------|-----------|-----------|-----------|
olumn Total |      1667 |      1666 |      3333 |
             |     0.500 |     0.500 |           |
-------------|-----------|-----------|-----------|


tatistics for All Table Factors


earson's Chi-squared test
------------------------------------------------------------
hi^2 =  24.62856     d.f. =  1     p =  6.951394e-07
```

روش equal-width binning:

```
(range = (max(as.numeric(new_data_set[, 1])) -
        min(as.numeric(new_data_set[, 1])) + 1) / numOfBins);

equal_length_data_set = new_data_set;

for(i in 1:numOfBins){
  for(j in 1:numOfRecords){
    if((i - 1) * range <= as.numeric(new_data_set[j, 1]) &&
       as.numeric(new_data_set[j, 1]) <= i * range){
      equal_length_data_set[j, 1] = level[i]
    }
  }
}

library(gmodels)
CrossTable(equal_length_data_set[, 2], equal_length_data_set[, 1],
           prop.r = FALSE, prop.t = FALSE
           , chisq = TRUE, prop.chisq = FALSE,
           dnn = c("Churn", "Day.Mins"))
```

```
42:47  (Untitled) :

Console  Terminal  Background Jobs

R 4.2.2 · ~/

             | Day.Mins
       Churn |      high |       low | Row Total |
-------------|-----------|-----------|-----------|
      False. |      1454 |      1396 |      2850 |
             |     0.830 |     0.882 |           |
-------------|-----------|-----------|-----------|
       True. |       297 |       186 |       483 |
             |     0.170 |     0.118 |           |
-------------|-----------|-----------|-----------|
Column Total |      1751 |      1582 |      3333 |
             |     0.525 |     0.475 |           |
-------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  18.16722     d.f. =  1     p =  2.023317e-05
```

در این روش نیز همانطور که دیده میشود، تفاوت درصد میان churn = True و churn = False در مقایسه سطری به طرز قابل توجهی وجود ندارد، و درصد های churn در گروه های low و high یکسان است.

امیرمهدی انصاری پور 810198358 سوالات فصل 3

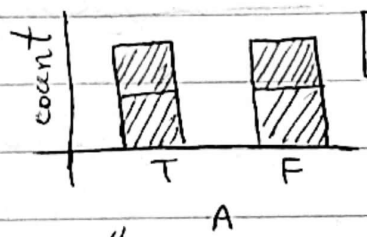
6) نمودار ترامبلی جدول توافقی چیست؟ clustered bar chart

7) با یک مثال، اثر متقابل (interaction) بین دو متغیر کیفی را بیان کنید.

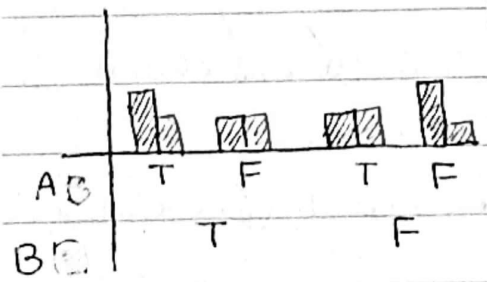
زمانیکه دو متغیر کیفی (A و B) بصورت همزمان رابطه شان با متغیر هدف بررسی میشود، در صورتیکه رابطه جدید بدست آمده با رابطه هرکدام از متغیرهای A, B متفاوت باشد، گوییم A و B با هم interaction دارند.

بصورت کلی با متغیر هدف قابل توجهی داشت، می گوییم A و B با هم تعادت دارند.

بعنوان مثال، فرض کنید رابطه متغیر A با دو مقدار T و F را با متغیر هدف churn بدست آوریم.



با توجه به barchart بالا، گویا متغیر مستقل A ارتباطی با target ندارد. اما حال متغیر B را در نظر می گیریم،



همانطور که دیده میشود، در دسته بندی جدید، در هرکدام از چهار حالت (A و B)، بین churn کشدان و آنهایی که churn نکردند تعادت معناداری وجود دارد. در نتیجه متغیر A با هم B و interaction دارند.

8) overlay histogram ... رابطه میان متغیر هدف و یک متغیر می تواند

• numerical را نشان دهد ... در هر ستون

برای نشان دادن proportion متغیر هدف

به طور واضح از حالت normalized استفاده می کنیم

9) مزیت normalized histogram : نسبت متغیر هدف را در هر ستون متغیر

پیشرو بطور واضح نشان میدهد و میتوان ارتباط میان متغیر هدف و پیشرو را

راحت تر و visible مشاهده کرد. (برای مثال با زیاد شدن مقدار متغیر پیشرو ،

درصد متغیر هدف churn

... نیز بیشتر میشود)

عیب normalized histogram : مورد فراوانی متغیر پیشرو اطلاعاتی نمی دهد. برای

مثال بیشتر تنها دونفر در یک ستون از histogram هستند که اگر آن دو

نفر churn شده، کل ستون در حالت normalized ، churn میشود و بیشتر

برداشت درستی از نمودار انجام نشود

با توجه به توضیحات داده شده، بهتر است برای overlay histogram ، هم نمودار

normalized ، هم non normalized را رسم کرد.

13) روش Binning براساس متغیر پیوسته: در این روش متغیر یعنی A پیوسته را براساس دسته متغیر هدف (دایجست) به تعدادی گروه تقسیم می‌کنیم و x (تبدیل به متغیر) می‌کنیم. برای مثال، متغیر A مقادیری بین 0 تا 10 می‌گیرد اما بیشتر داده‌ها در churn recordهای متغیر A با مقدار بیشتر از 3، چنین برای recordهای با مقدار کمتر از 3 است. بنابراین، می‌توان متغیر A را به دو گروه low (کمتر از 3) و high (بیشتر از 3) قرار داد.

15) باید قبل از derive کردن متغیر جدید به scale متغیرهای numerical قبلی توجه کنیم. برای مثال، اگر متغیر عددی A، میانگین و انحراف معیار بیشتری از متغیر B دارد، میانگین ترکیب از recordهای A و B باعث بیشتر شدن bias نسبت به متغیر A می‌شود. پس بهتر است قبل از میانگین ترکیب، هر دو متغیر A و B را به Z_score شان تبدیل کنیم.

19) میان هر دو متغیرهای پیوسته عددی، Pearson correlation را بدست می‌آوریم. چنانچه این مقدار ±1 بود، یکی از دو متغیر پیوسته را حذف می‌کنیم (این کار را میتوان با matrixplot و یا با جدول correlation انجام داد) برای متغیرهای پیوسته زوج

با correlation بالا و p-value کم باید برای ماژول بعدی Data Mining کاندید برای حذف شوند.

(20

Bar charts: i: رسته‌ای (یعنی)

Histograms: عددی

Summary statistics: عددی (برای barplot ها و مدوس را نشان میدهد)

Crosstabulations: رسته‌ای

Correlation analysis: عددی (پیوسته)

Scatter plot: عددی

Binning: عددی