Introduccion a Data Mining
Trabajo Práctico

Churn - Telco

Churn, también llamado *attrition*, es un término usado para indicar un cliente que deja el servicio de una compañía en favor de otra compañía.

El data set contiene 20 variables con datos sobre ~3000 clientes, junto con una indicación de si el cliente dejó (churned) o no la compañía.

Churn Data set. Blake, C.L. & Merz, C.J. UCI Repository of machine learning databases [kdd.ics.uci.edu/]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

Las variables son las siguientes:

1 _ *State:* categorical, for the 50 states and the District of Columbia
2 _ *Account length:* integer-valued, how long account has been active
3 _ *Area code:* categorical
4 _ *Phone number:* essentially a surrogate for customer ID
5 _ *International Plan:* dichotomous categorical, yes or no
6 _ *VoiceMail Plan:* dichotomous categorical, yes or no
7 _ *Number of voice mail messages:* integer-valued
8 _ *Total day minutes:* continuous, minutes customer used service during the day
9 _ *Total day calls:* integer-valued
10_ *Total day charge:* continuous, perhaps based on foregoing two variables
11_ *Total evening minutes:* continuous, minutes customer used service during the evening
12_ *Total evening calls:* integer-valued
13_ *Total evening charge:* continuous, perhaps based on foregoing two variables
14_ *Total night minutes:* continuous, minutes customer used service during the night
15_ *Total night calls:* integer-valued
16_ *Total night charge:* continuous, perhaps based on foregoing two variables
17_ *Total international minutes:* continuous, minutes customer used service to make international calls
18_ *Total international calls:* integer-valued
19_ *Total international charge:* continuous, perhaps based on foregoing two variables
20_ *Number of calls to customer service:* integer-valued.

Análisis Descriptivo Previo del dataset *churn*

1. Explore si hay valores faltantes en alguna de las variables.
2. Compare los campos *area code* y *state*. Discuta cualquier aparente anormalidad.
3. Emplee gráficos para determinar visualmente si hay algún valor extremo en la cantidad de llamadas a la línea de atención al cliente.
4. Identifique el rango de llamadas a *customer service*, que debieran considerarse *outliers* empleando:
    a) El método de puntaje z
    b) El método del RIC.
5. Transforme la variable *day minutes* empleando estandarización por puntaje Z.
6. Trabaje con los sesgos:
    a) Calcule el sesgo de la variable *day minutes.*
    b) Calcule el sesgo de la variable estandarizada por puntaje Z para *day minutes*. Comente.
    c) Basado en el valor del sesgo, ¿Considera que la variable se encuentra sesgada o es casi perfectamente simétrica?
7. Construya el normal probability plot de la variable *day minutes.* Comente sobre la normalidad de los datos.

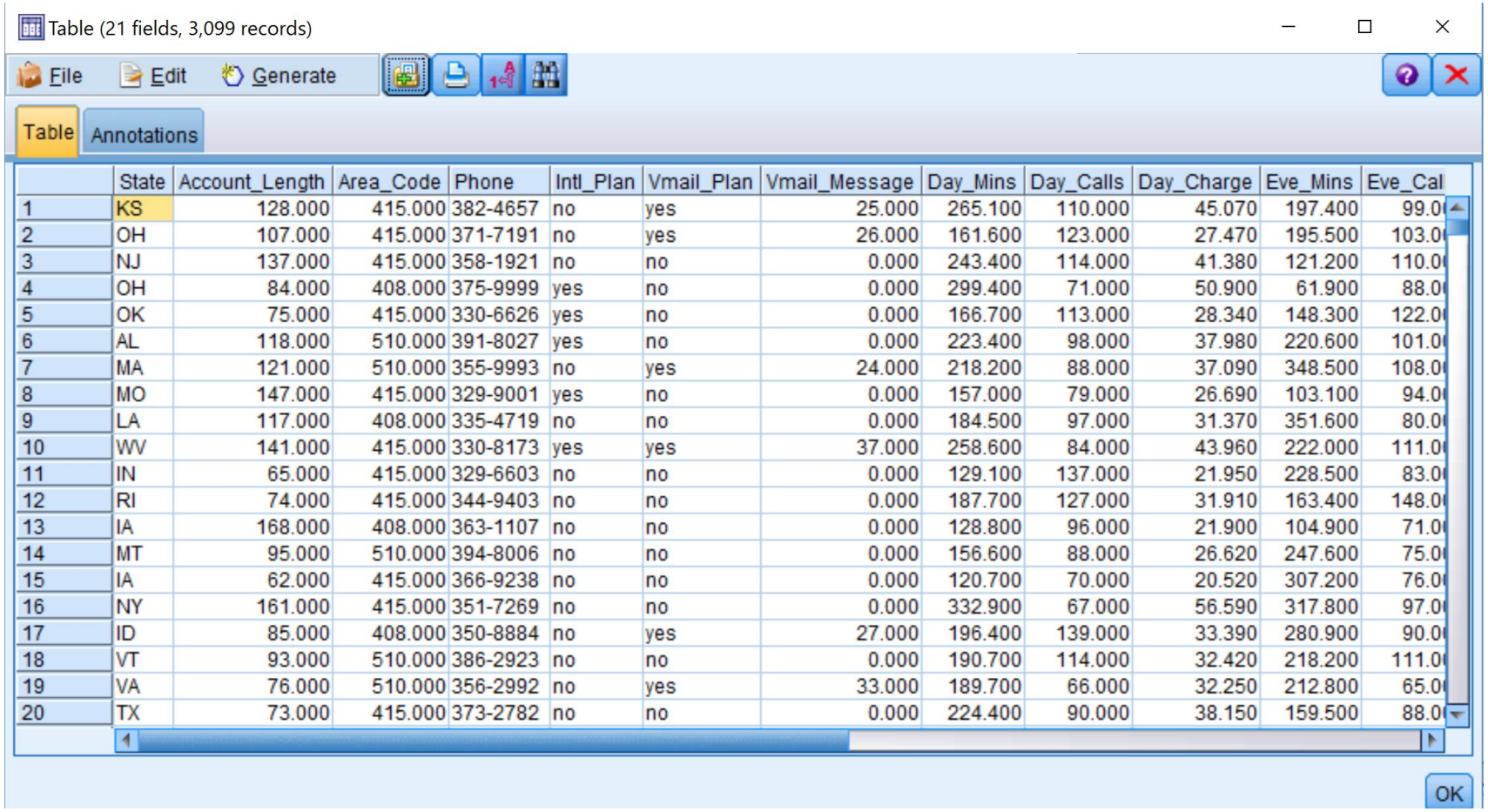
Análisis Descriptivo Previo del dataset *churn*

8. Trabaje con la variable *international minutes:*
    a) Construya el normal probability plot de la variable.
    b) ¿Que evita que esta variable tenga una distribución normal?
    c) Construya una variable indicadora para lidiar con la situación anterior.
    d) Construya un normal probability plot de la variable derivada *nonzero international minutes*. Comente en relación a la normalidad de la variable derivada.
9. Transforme la variable *night minutes* empleando estandarización por puntaje Z. Empleando un grafico, describa el rango de los valores estandarizados.

Caso Telco – Tareas a Realizar

- Observar las distribuciones de las variables
- Buscar y eliminar variables correlacionadas.
- Analizar las proporciones de churn para distintas variables.
- Particionar los datos.
- Generar tres modelos de árbol.
- Evaluar los modelos.

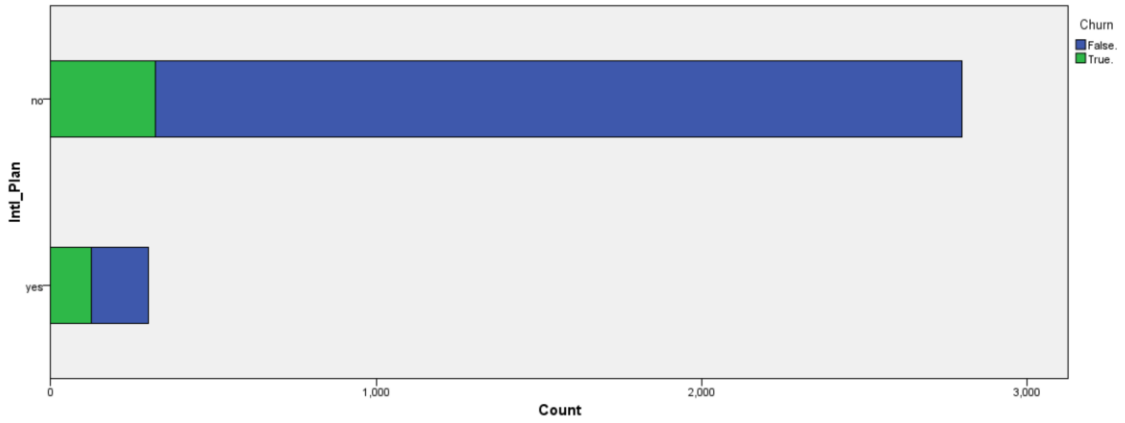
Ej. Modeler 17



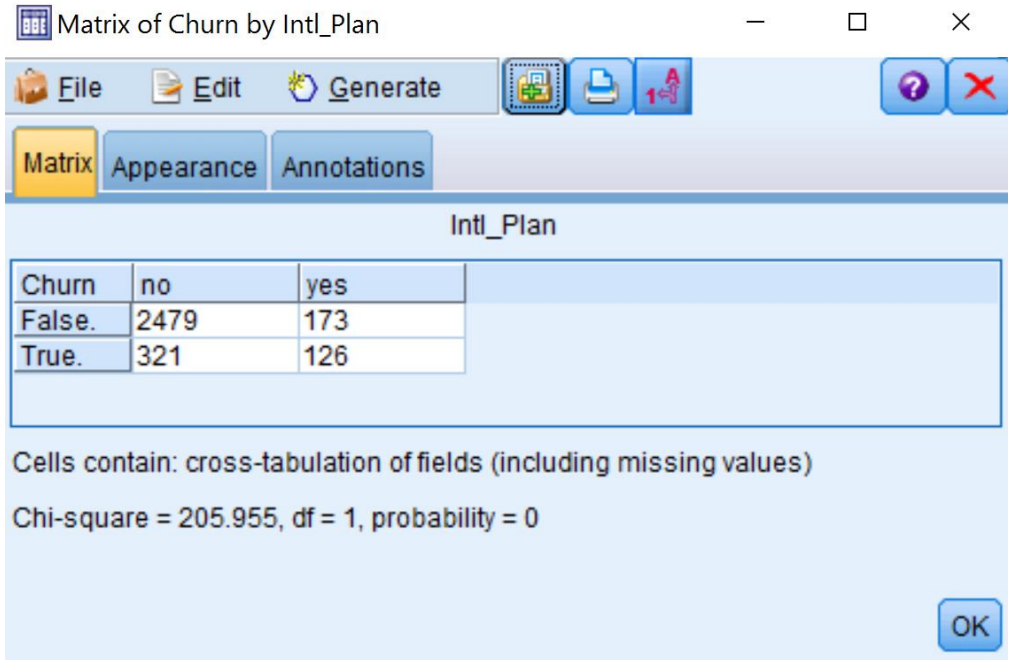
	State	Account_Length	Area_Code	Phone	Intl_Plan	Vmail_Plan	Vmail_Message	Day_Mins	Day_Calls	Day_Charge	Eve_Mins	Eve_Cal
1	KS	128.000	415.000	382-4657	no	yes	25.000	265.100	110.000	45.070	197.400	99.0
2	OH	107.000	415.000	371-7191	no	yes	26.000	161.600	123.000	27.470	195.500	103.0
3	NJ	137.000	415.000	358-1921	no	no	0.000	243.400	114.000	41.380	121.200	110.0
4	OH	84.000	408.000	375-9999	yes	no	0.000	299.400	71.000	50.900	61.900	88.0
5	OK	75.000	415.000	330-6626	yes	no	0.000	166.700	113.000	28.340	148.300	122.0
6	AL	118.000	510.000	391-8027	yes	no	0.000	223.400	98.000	37.980	220.600	101.0
7	MA	121.000	510.000	355-9993	no	yes	24.000	218.200	88.000	37.090	348.500	108.0
8	MO	147.000	415.000	329-9001	yes	no	0.000	157.000	79.000	26.690	103.100	94.0
9	LA	117.000	408.000	335-4719	no	no	0.000	184.500	97.000	31.370	351.600	80.0
10	WV	141.000	415.000	330-8173	yes	yes	37.000	258.600	84.000	43.960	222.000	111.0
11	IN	65.000	415.000	329-6603	no	no	0.000	129.100	137.000	21.950	228.500	83.0
12	RI	74.000	415.000	344-9403	no	no	0.000	187.700	127.000	31.910	163.400	148.0
13	IA	168.000	408.000	363-1107	no	no	0.000	128.800	96.000	21.900	104.900	71.0
14	MT	95.000	510.000	394-8006	no	no	0.000	156.600	88.000	26.620	247.600	75.0
15	IA	62.000	415.000	366-9238	no	no	0.000	120.700	70.000	20.520	307.200	76.0
16	NY	161.000	415.000	351-7269	no	no	0.000	332.900	67.000	56.590	317.800	97.0
17	ID	85.000	408.000	350-8884	no	yes	27.000	196.400	139.000	33.390	280.900	90.0
18	VT	93.000	510.000	386-2923	no	no	0.000	190.700	114.000	32.420	218.200	111.0
19	VA	76.000	510.000	356-2992	no	yes	33.000	189.700	66.000	32.250	212.800	65.0
20	TX	73.000	415.000	373-2782	no	no	0.000	224.400	90.000	38.150	159.500	88.0

- La variable *phone* emplea solamente siete (7) dígitos.
- Hay dos variables indicadoras.
- La mayoría de las variables son continuas.
- La variable respuesta *churn* es una variable indicadora con dos valores, *TRUE y FALSE*.

Exploración de Variables Categóricas



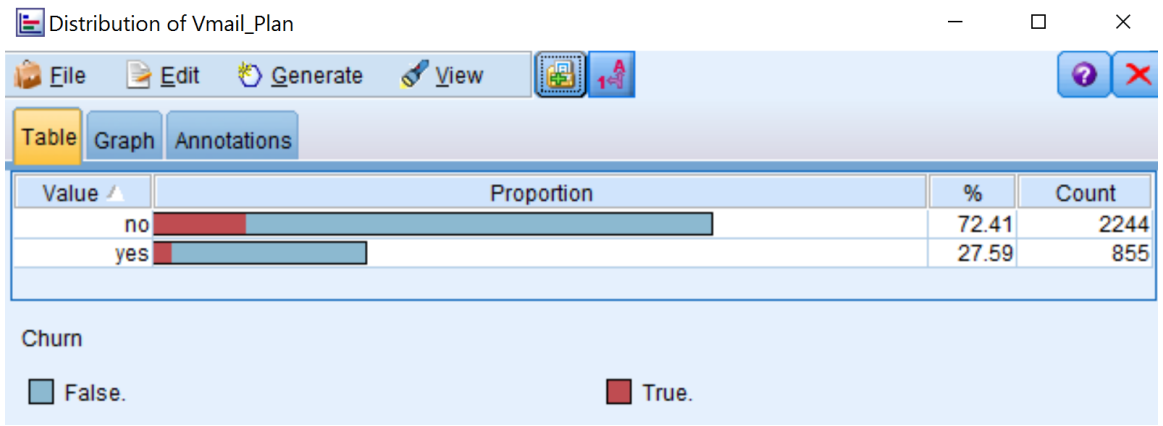
Proporción de churnes (verde) y no-churners en relación con tenencia de Plan Internacional o no. El 9,65 % de los clientes selecciono el Plan internacional, y este grafico parece indicar que una mayor proporción de los que ha seleccionado dicho plan esta abandonando la compañía.

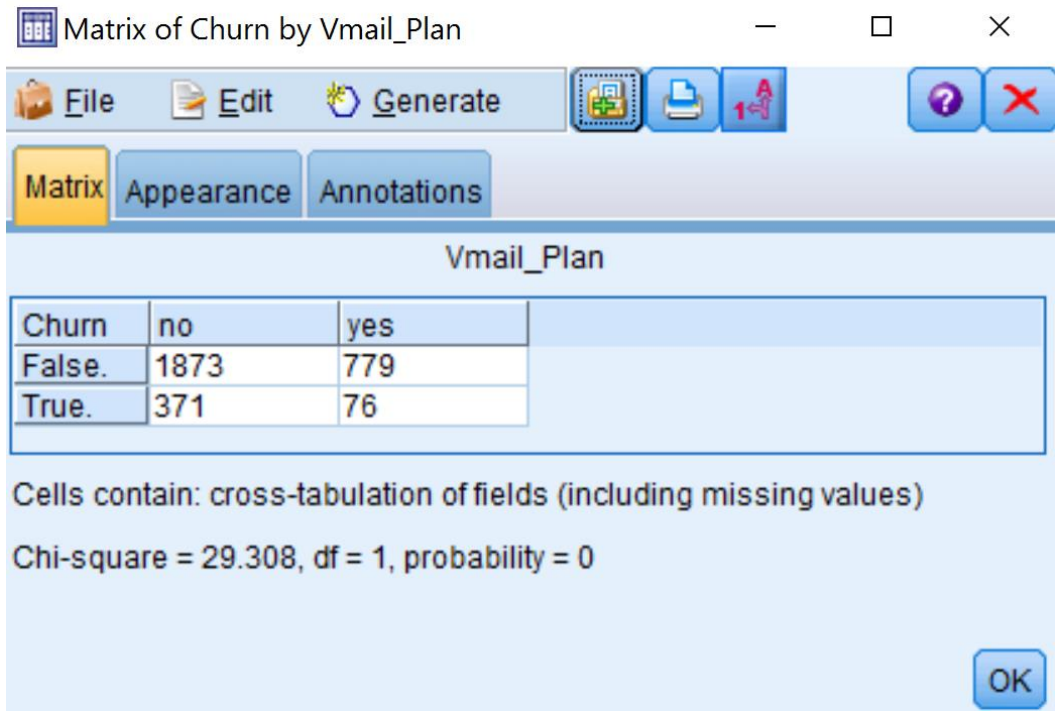


Churn x Intl_Plan
2479 + 321 = 2800 no eligieron el Plan Internacional
173 + 126 = 299 eligieron el Plan Internacional
La cantidad de churners es de 447, (447 / (447+2652) = 14, 4 %)
126 / (126 + 173) = 41 % de los clientes c/Plan Internacional son churners, comparados contra el 321 / (2479+321) = 11 % de los clientes s/Plan Internacional que son churners…

Algunas conclusiones preliminares

- Debemos investigar un poco mas en relación que hace en el Plan Internacional a los clientes abandonar la compañía.
- Probablemente cualquier algoritmo de *data mining* que empleemos para predecir *churn*, incluirá la variable sobre elección del Plan Internacional de llamadas.





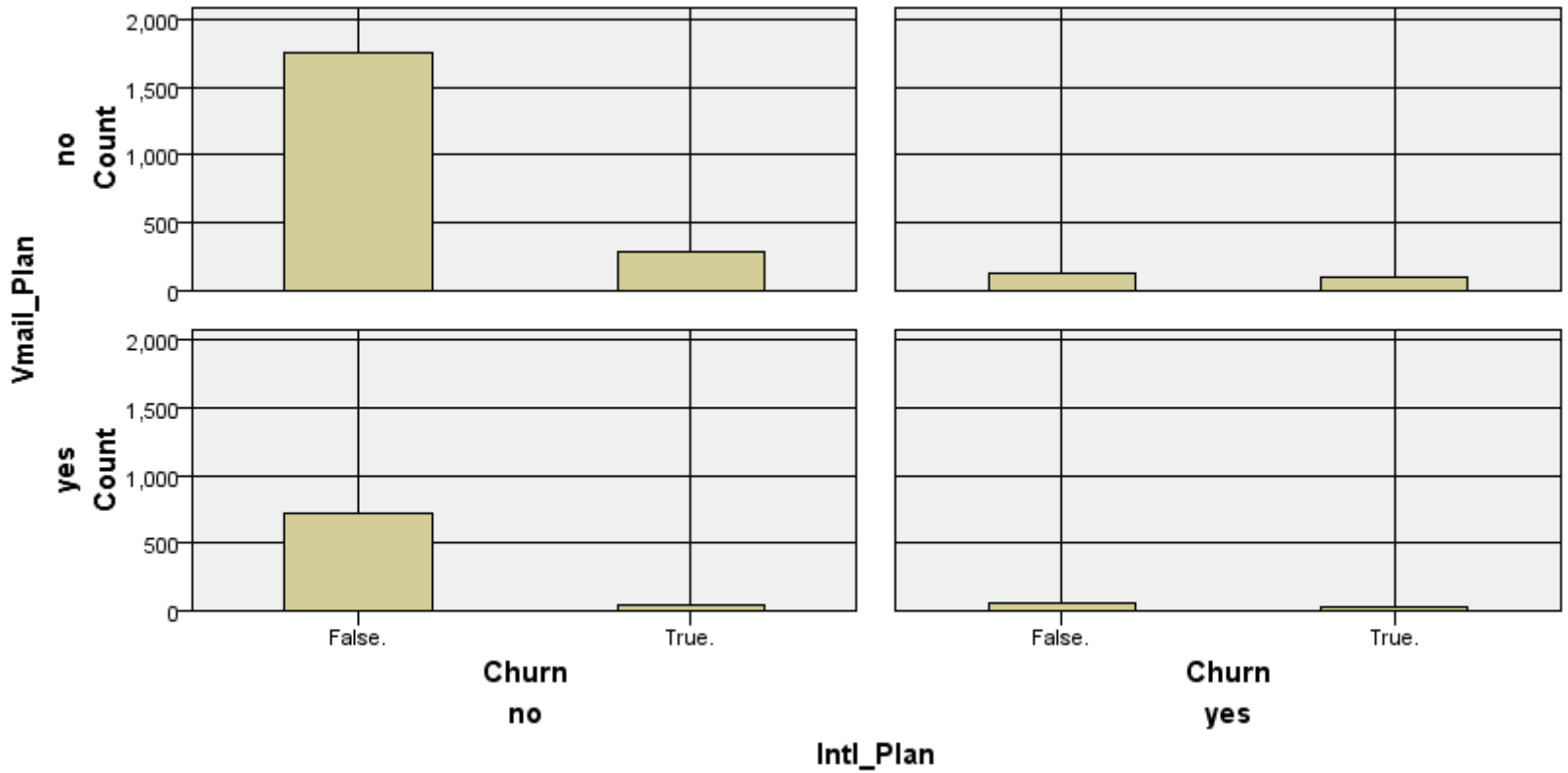
Voice_Mail_Plan = 779 + 76 = 855
Sin Voice Mail_Plan = 1873 + 371 = 2244

Churners = 371 + 76 = 447
No churners = 2652

371 / 2244 = 16,5 % de los que NO tienen Voice_Mail son churners, mientras que 8,9 % de los que SI lo tienen son churners. Es decir, los clientes que no tienen el Voice Mail son dos veces mas probable que abandonen la compañía.

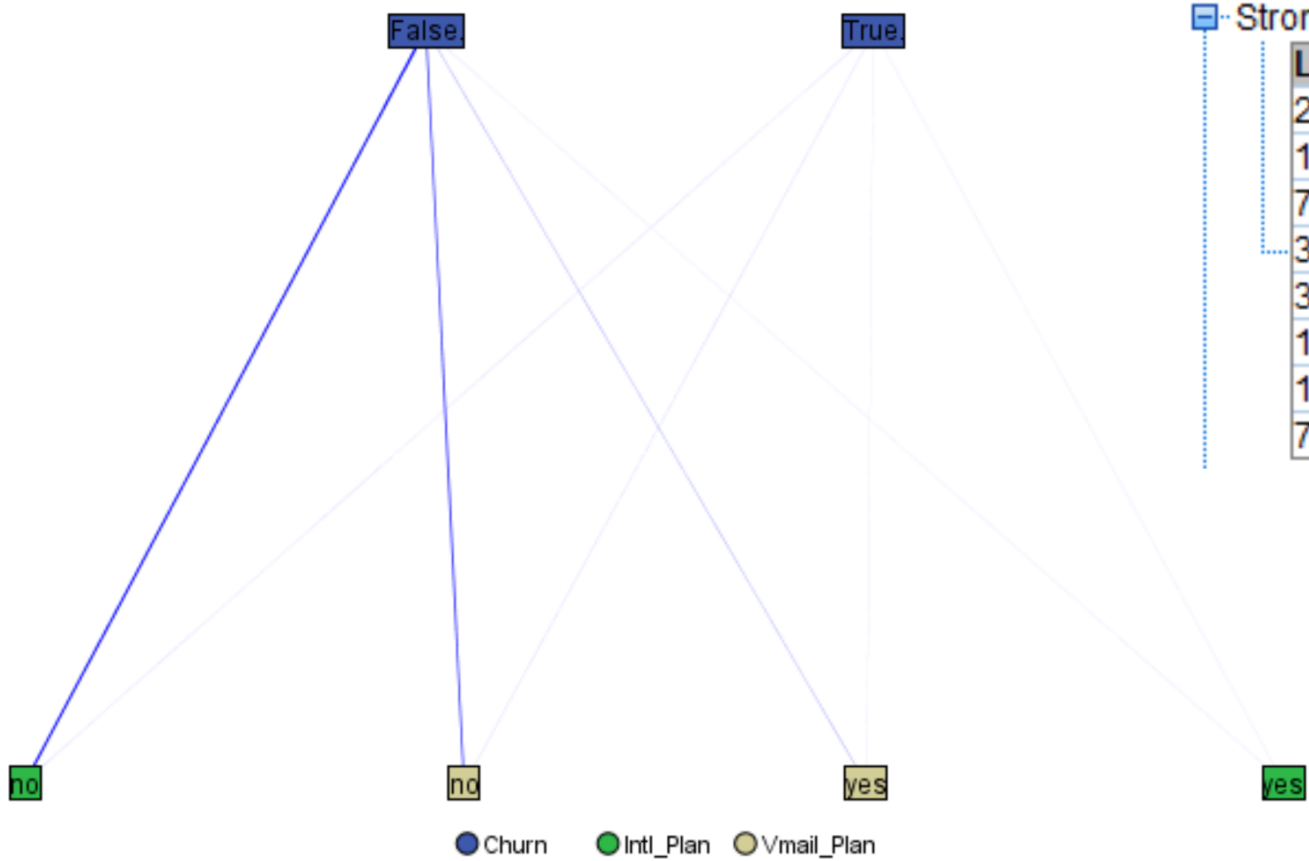
Algunas conclusiones preliminares

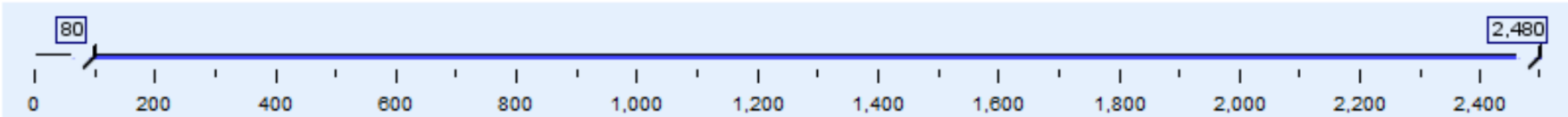
- Debemos investigar un poco mas en relación que hace en el Plan Voice Mail para retener a los clientes de la compañía.
- Probablemente cualquier algoritmo de *data mining* que empleemos para predecir *churn*, incluirá la variable sobre elección del Plan Voice Mail de llamadas.



- Voice_Mail NO:
  - Hay mas clientes que no tienen ningún plan que aquellos que tienen solo el plan internacional.
  - En los clientes que no tienen Voice_Mail la proporción de churners es mayor si tienen Internacional Plan (44 % que si no lo tienen 14 %)
- Voice_Mail SI
  - Hay mas clientes que tienen Voice_Mail que aquellos que tienen las dos opciones
  - La proporción de chuners es mayor si tienen Plan Internacional (39 % a 5 %)

Directed Web of Web



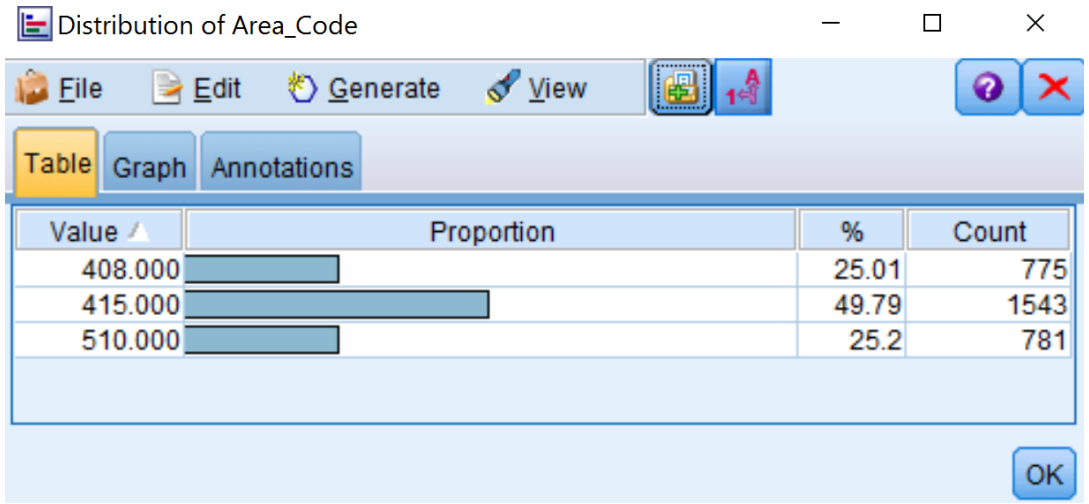


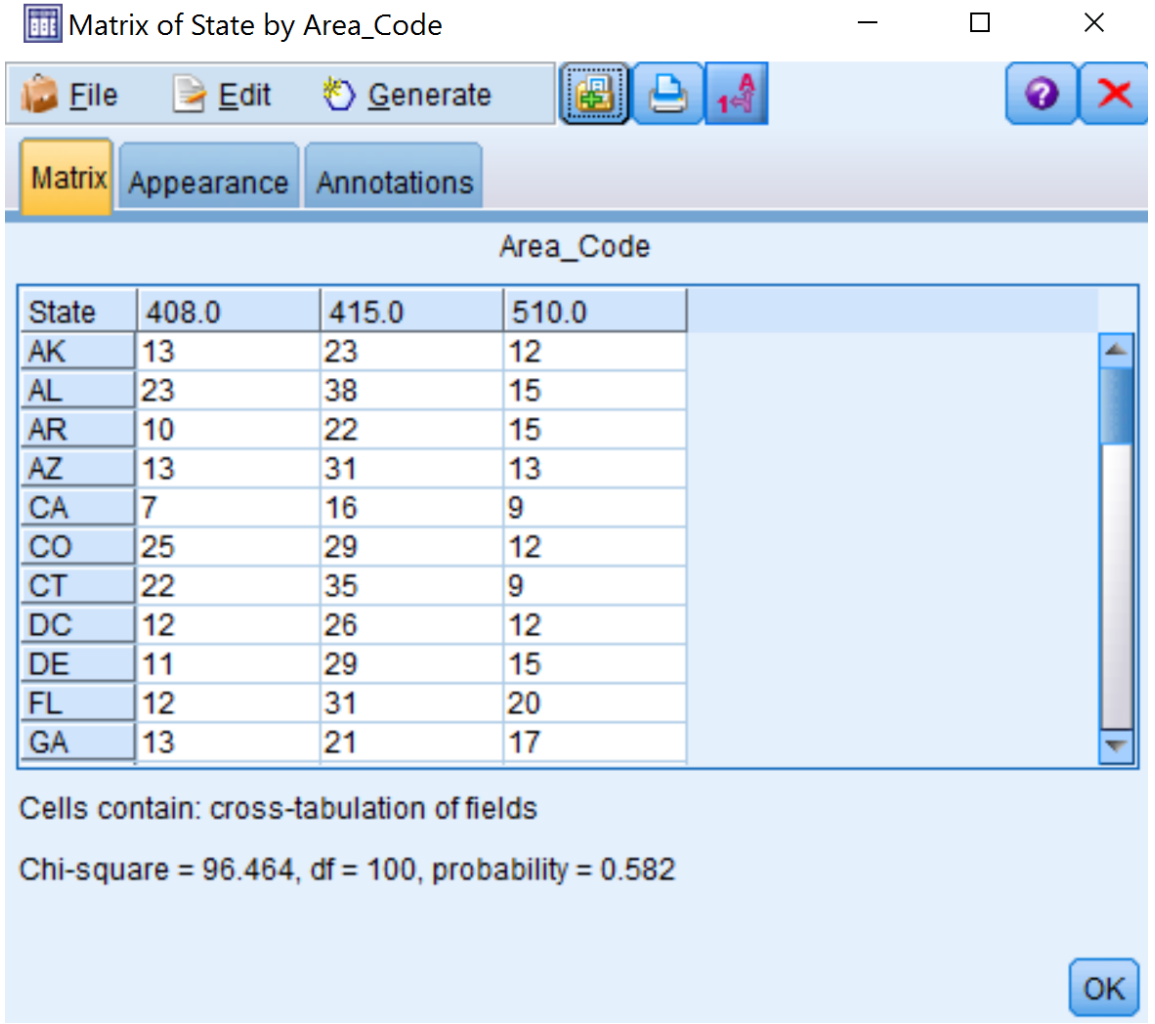
Strong Links

Links	Field 1	Field 2
2,479	Intl_Plan = "no"	Churn = "False."
1,873	Vmail_Plan = "no"	Churn = "False."
779	Vmail_Plan = "yes"	Churn = "False."
371	Vmail_Plan = "no"	Churn = "True."
321	Intl_Plan = "no"	Churn = "True."
173	Intl_Plan = "yes"	Churn = "False."
126	Intl_Plan = "yes"	Churn = "True."
76	Vmail_Plan = "yes"	Churn = "True."

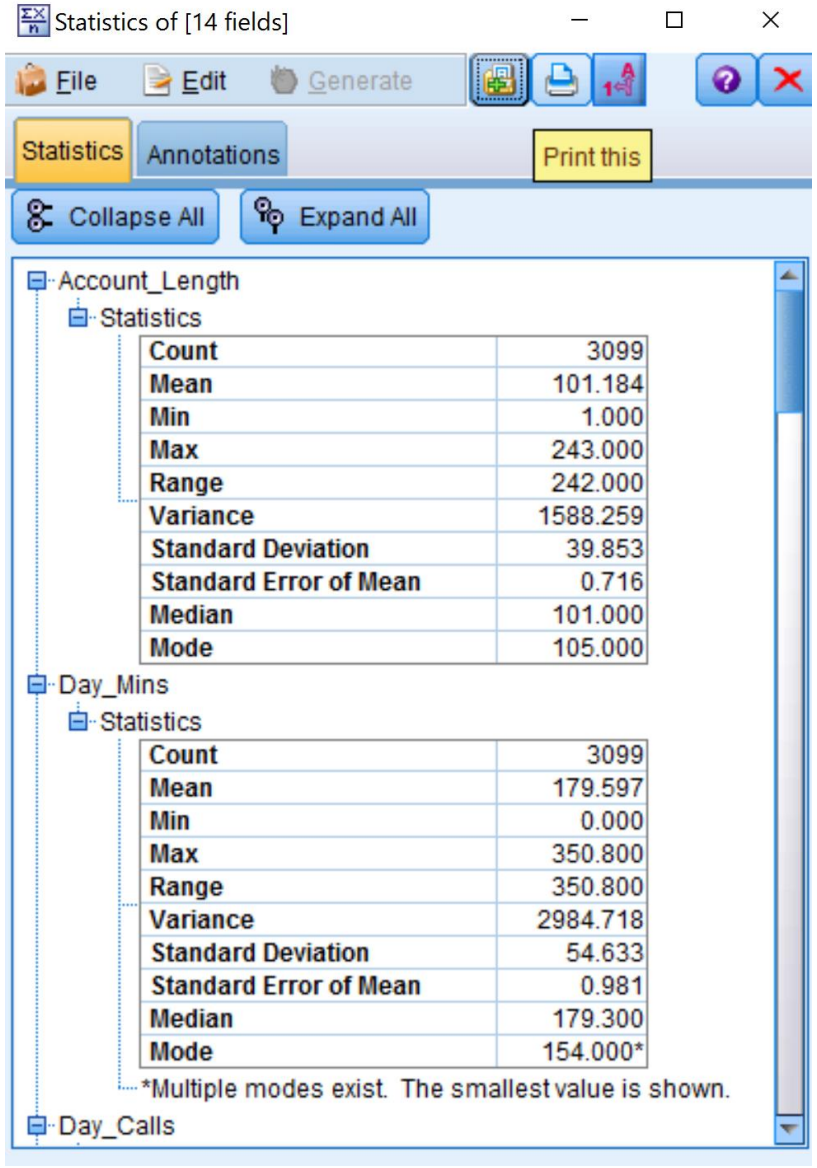
Campos anómalos?

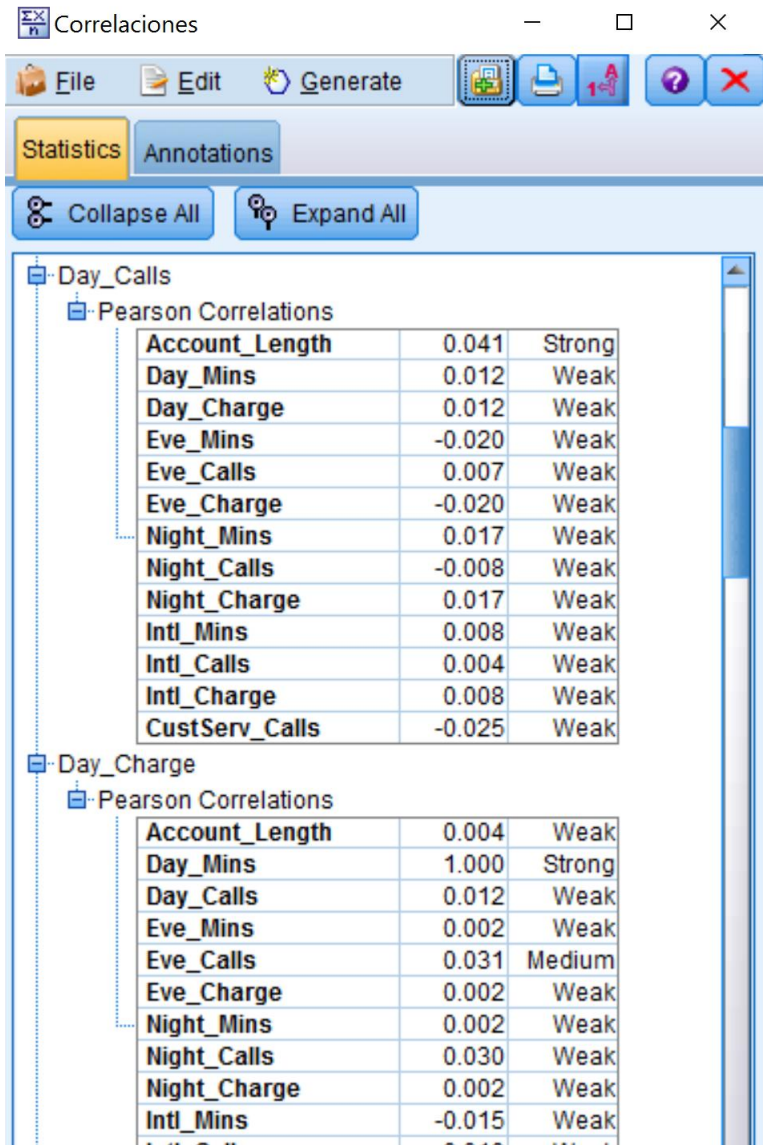
- Analizar Area_Code y State.

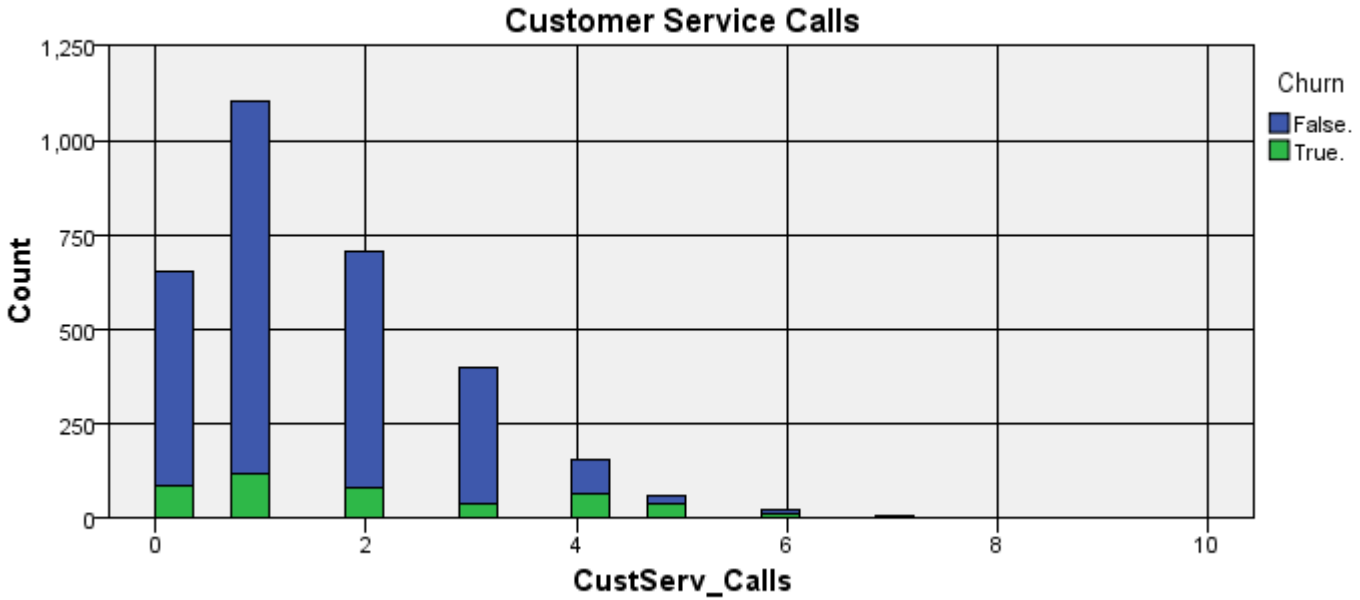




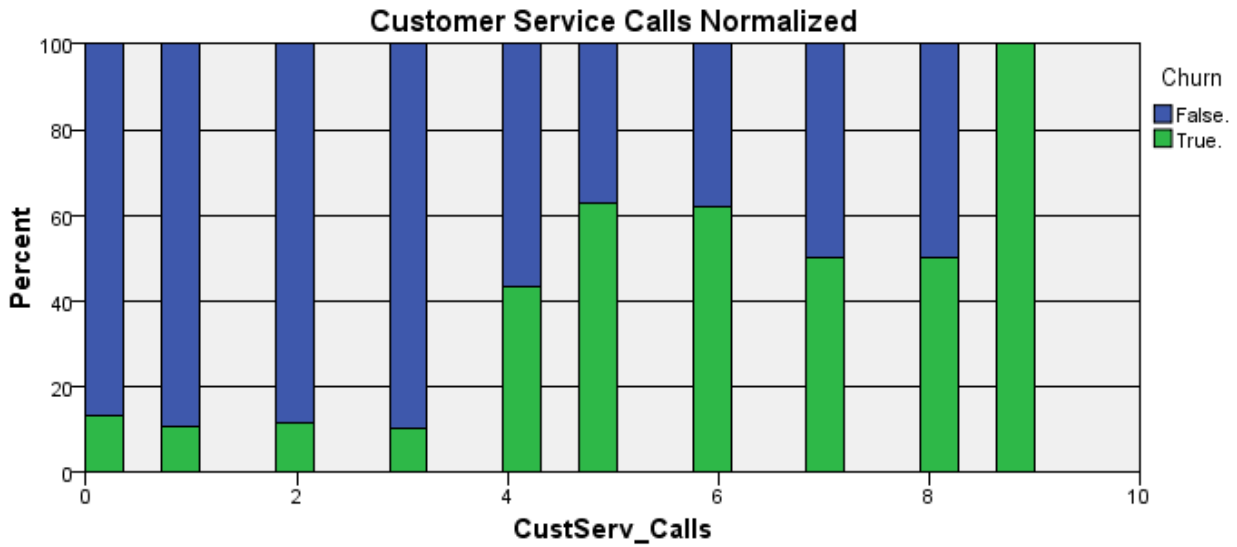
Exploración de variables numéricas

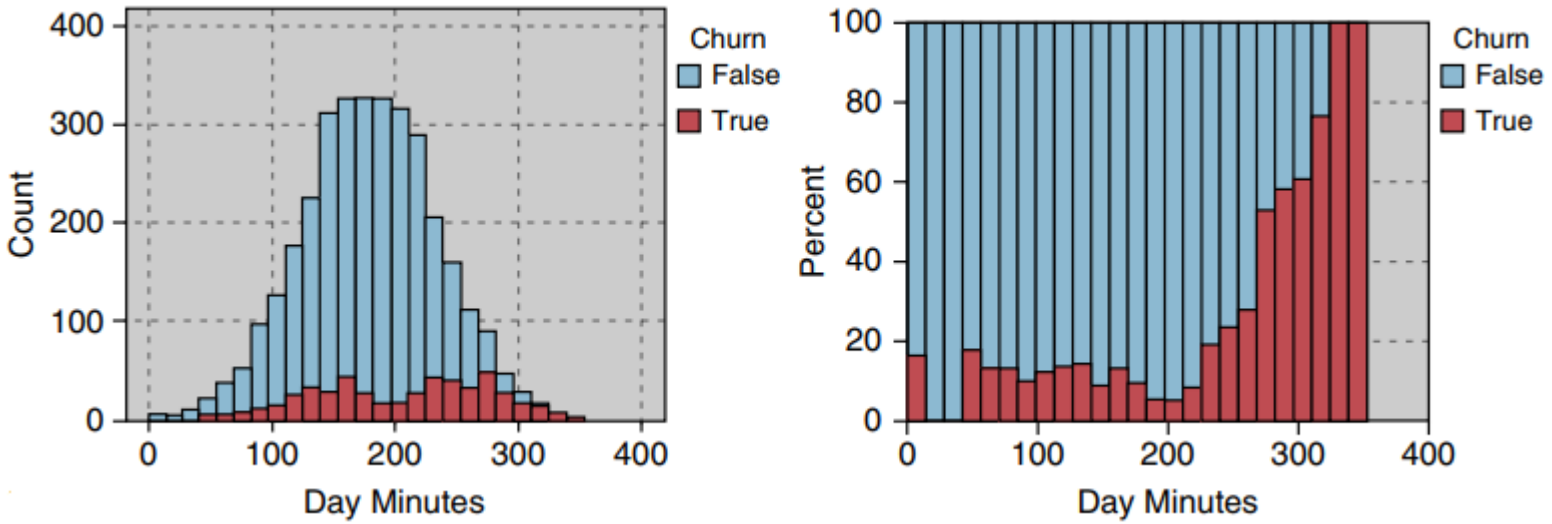






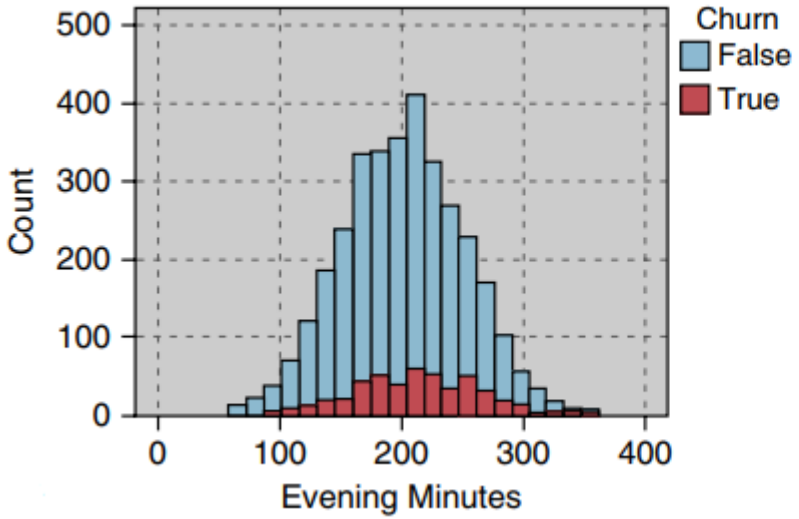
- Es necesario controlar la cantidad de llamadas a servicio al cliente. Luego de la tercer llamada, es necesario ofrecer incentivos para retener al cliente.
- Esperamos que cualquier algoritmo de data mining que empleemos para predecir el *churn,* emplee esta variable entre sus predictores.

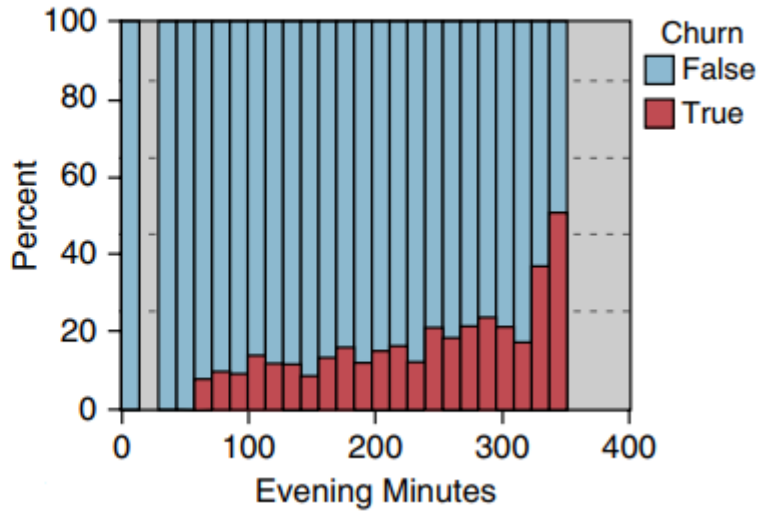




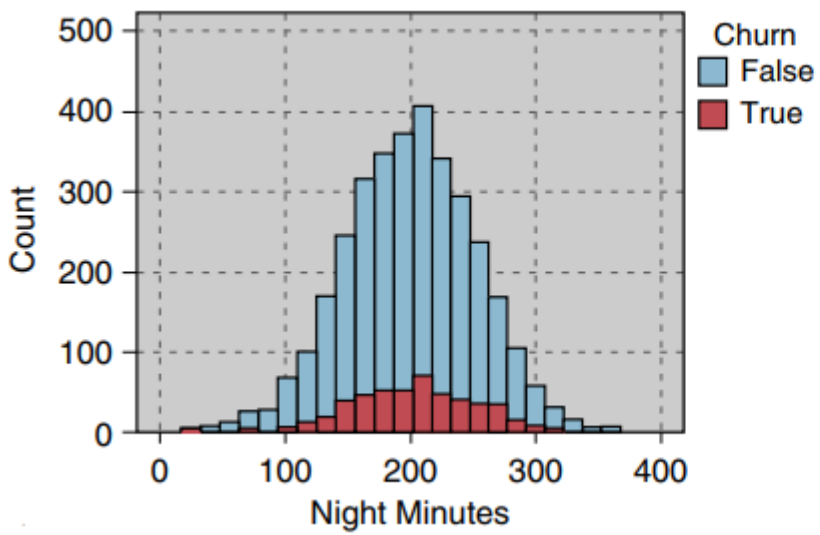
Los usuarios con valores altos en *day minutes* tienden a irse a tasas mas altas:

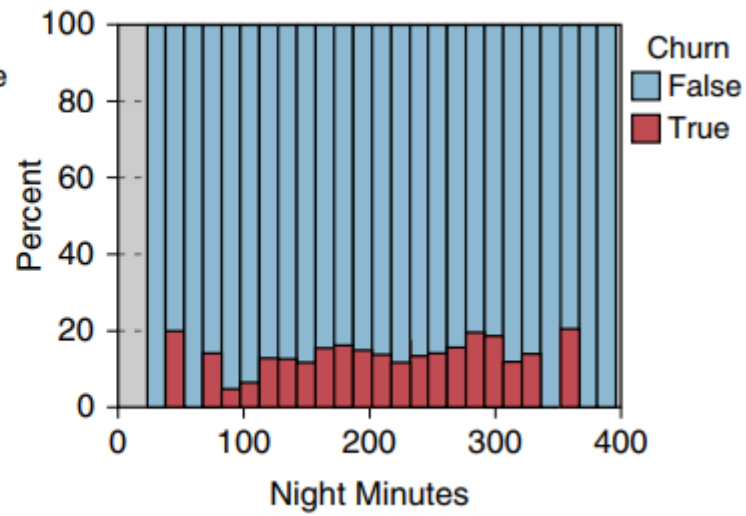
1. Podríamos controlar cuidadosamente la cantidad de minutos utilizada por el cliente, cuando el valor supera 200, debiéramos considerar incentivos especiales
2. Debiéramos investigar el motivo por lo que los usuarios intensivos de minutos de día están tentados a irse.
3. Esperaríamos que nuestro modelo de *data mining*, incluya *day minutes* como un predictor de churn.

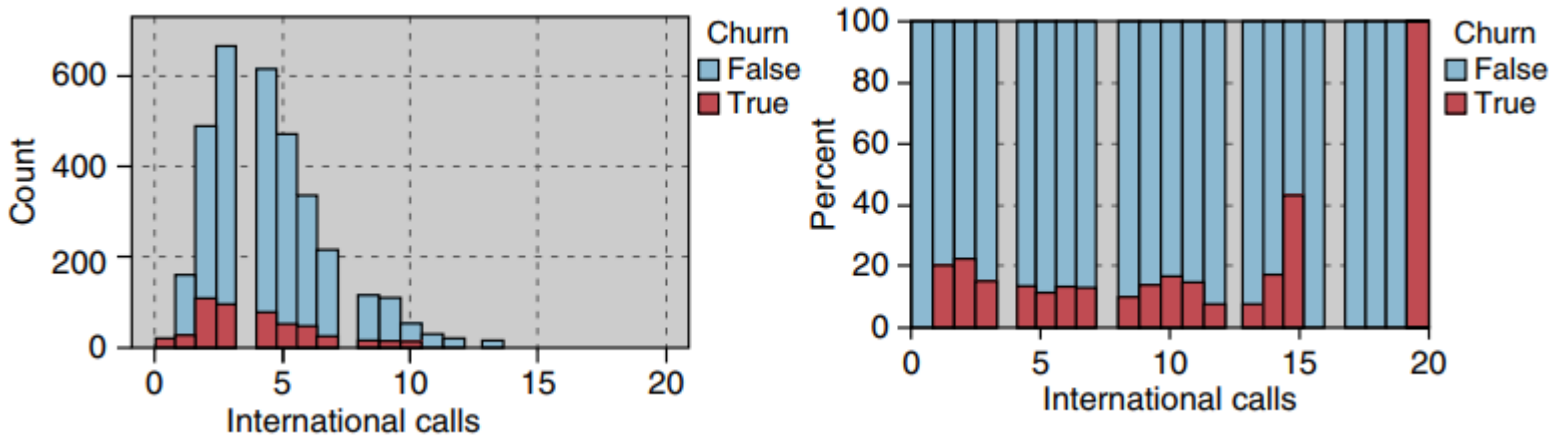




Los usuarios con valores altos en *day minutes* también podrían tender a irse a tasas mas altas, pero no es tan claro como el caso anterior.







Estos graficos no parecen indicar una fuerte evidencia de la importancia predictiva de la variable *international Calls*. Pero un test t- para la diferencia en la media de las llamadas internacionales para los churners y los no churners es estadísticamente signficativo, indicando que la variable efectivamente es útil para predecir el *churn.*

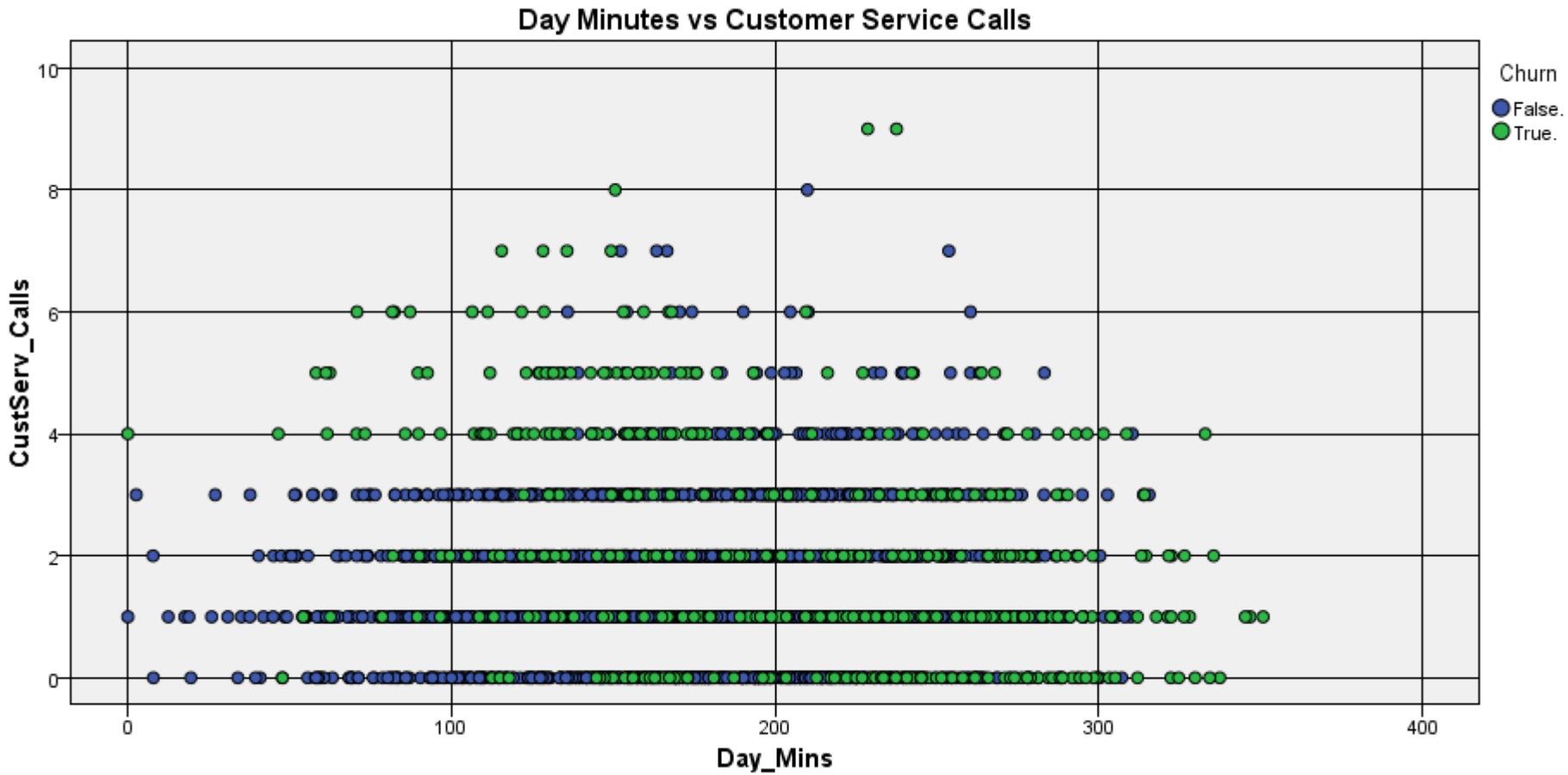
Los Churners tienden a colocar menor cantidad promedio de llamadas internacionales…

Two-Sample T-Test and CI: Intl Calls, Churn

Two-sample T for Intl Calls

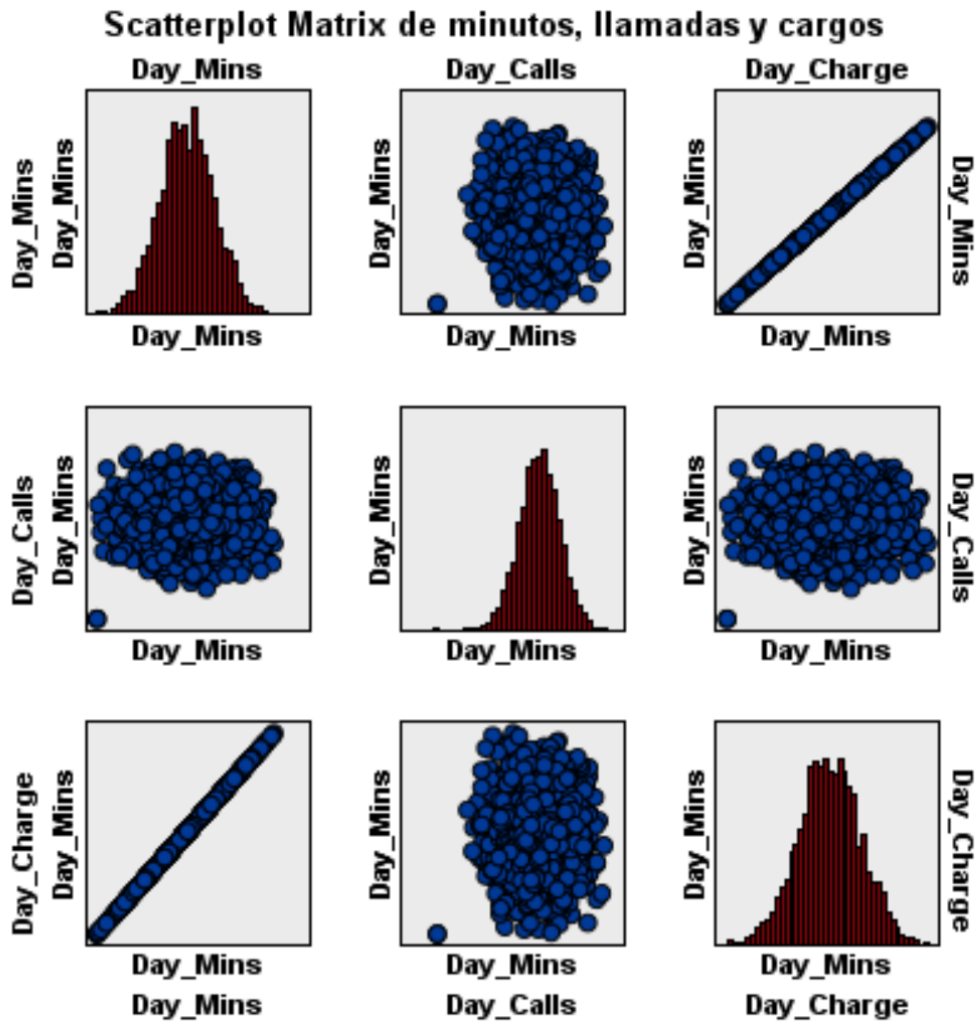
Churn	N	Mean	StDev	SE Mean
False	2850	4.53	2.44	0.046
True	483	4.16	2.55	0.12

Difference = mu (False) - mu (True)
Estimate for difference: 0.369
95% CI for difference: (0.124, 0.614)
T-Test of difference = 0 (vs not =): T-Value = 2.96 **P-Value = 0.003** DF = 640


Day Minutes vs Customer Service Calls
CustServ_Calls
Day_Mins
Churn
False.
True.
0
2
4
6
8
10
0
100
200
300
400

Hay variables correlacionadas?

- *charge* podría ser función de *minutes* y de *llamadas*…



- No parece haber relación entre *day_minutes* y *day_calls* o entre *day_calls* y *day_charge.*
(esto puede parecer raro, ya que *a priori* hubiéramos asumido que a medida que aumenta la cantidad de llamadas, la cantidad de minutos también tendería a aumentar (y lo mismo para la facturación), es decir una correlación positiva entre esas variables, pero eso no se aprecia gráficamente y tampoco con el calculo de la correlación r = 0.07 para ambos casos)

- Pero parece existir una correlación lineal perfecta entre *day_minutes* y *day_charge*, indicando que *day_charge* es una función lineal solamente de *day_minutes.*

Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	Day_Mins[b]	.	Enter

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000[a]	1.000	1.000	.002862

a. Predictors: (Constant), Day_Mins

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	267226.798	1	267226.798	3.263E+10	.000[b]
	Residual	.025	3097	.000		
	Total	267226.823	3098			

b. Predictors: (Constant), Day_Mins

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.001	.000		3.634	.000
	Day_Mins	.170	.000	1.000	180628.382	.000

day_charge = 0.000613 + 0.17**day_minutes*

(modelo de tarifa plana, 17 centavos por minuto)

Dado que *day_charge* y *day_minutes* están perfectamente correlacionadas, eliminamos una de ellas.
Investigar *evening, night* e *international*

Resumen del Análisis Exploratorio

- Los cuatro campos *charge* son funciones lineales de los campos *minutos*, y por lo tanto debieran ser omitidos del análisis.
- El campo *area code* y/o el campo *state* son anómalos, y por lo tanto debieran ser omitidos hasta obtener alguna clarificación al respecto.
- En relación con *churn:*
  - Los clientes con *International Plan* tienden a irse con mayor frecuencia.
  - Los clientes con *Voice Mail Plan* tienden a irse con menor frecuencia.
  - Los clientes con cuatro o mas *Customer Service Calls* se van mas de cuatro veces mas frecuentemente que los demás clientes.
  - Los clientes con altos *day minutes* y altos *evening minutes* tienden a irse a una mayor tasa que los demás clientes. (alrededor de seis veces mayor que los demás clientes).
  - Los clientes con bajo *day minutes* y altos *customer service calls* se van a una tasa mas alta que los demás clientes.
  - Los clientes con baja cantidad de *International Calls* se van a una tasa mas alta que los clientes con mayor cantidad de llamadas internacionales.

MODELOS PREDICTIVOS?