Exploratory Data Analysis (EDA)

HYPOTHESIS TESTING VERSUS EXPLORATORY DATA ANALYSIS

- Hypotheses tests relationships between variables.
- E.g. Cell-phone executives are interested in whether a recent increase in the fee structure has led to a decrease in market share.
- Many statistical hypothesis testing procedures are available.
- Especially when confronted with unknown, large databases, analysts often prefer to use Exploratory Data Analysis (EDA), or graphical data analysis.

Exploratory Data Analysis

- **Exploratory Data Analysis (EDA)** is that part of statistical practice concerned with reviewing, communicating and using data where there is a low level of knowledge about its cause system.
- Many **EDA techniques** have been adopted into data mining and are being taught to young students **as a way to introduce them to statistical thinking**.

- **www.wikipedia.org**

Objectives of EDA

EDA allows the analyst to-

- delve into the data set;
- examine interrelationships among attributes;
- identify interesting subsets of the observations;
- develop an initial idea of possible associations amongst the predictors, as well as between the predictors and the target variable.

HYPOTHESIS TESTING VERSUS EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly **graphical** and **statistical**) to maximize

1. insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop accurate models;

GETTING TO KNOW THE DATA SET

- Graphs, plots, and tables often uncover important relationships.
- Relationships that could indicate important areas for further investigation.
- We will use exploratory methods to delve into the **churn data set** from the **UCI Repository of Machine Learning Databases** at the University of California
- Churn, also called attrition, is a term used to indicate a customer leaving the service of company.

Churn data set

The data set contains **20 predictors**.

- *State*: Categorical, for the 50 states and the District of Columbia.
- *Account length:* Integer-valued, how long account has been active.
- *Area code:* Categorical
- *Phone number*: Essentially a surrogate for customer ID.
- *International plan*: categorical, yes or no.
- *Voice mail plan:* categorical, yes or no.
- *Number of voice mail messages:* Integer-valued.
- *Total day minutes:* Continuous, minutes customer used service during the day.
- *Total day calls:* Integer-valued.
- *Total day charge:* Continuous, perhaps based on above two variables.

Churn data set

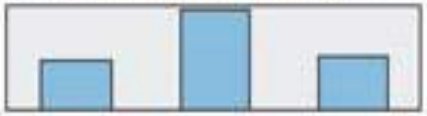
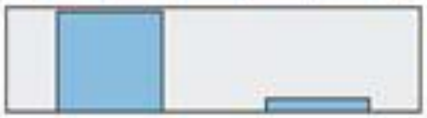
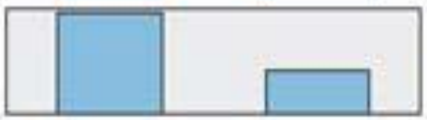
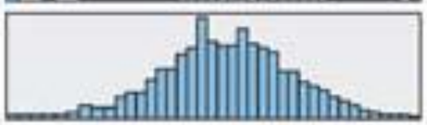
- *Total eve minutes*: Continuous, minutes customer used service during the evening.
- *Total eve calls*: Integer-valued.
- *Total eve charge*: Continuous, based on above two variables.
- *Total night minutes*: Continuous, minutes customer used service during the night.
- *Total night calls*: Integer-valued.
- *Total night charge*: Continuous, perhaps based on above two variables.
- *Total international minutes*: Continuous, minutes customer used service to make international calls.
- *Total international calls*: Integer-valued.
- *Total international charge*: Continuous, based on above two variables.
- *Number of calls to customer service*: Integer-valued.
- ***Churn***: **Target**. Indicator of whether customer has left company **(true or false).**

Field values of the first 10 records in the *churn* data set

	State	Account Length	Area Code	Phone	Intl Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins
1	KS	128	415	382-4657	no	yes	25	265.100	110	45.070	197.400
2	OH	107	415	371-7191	no	yes	26	161.600	123	27.470	195.500
3	NJ	137	415	358-1921	no	no	0	243.400	114	41.380	121.200
4	OH	84	408	375-9999	yes	no	0	299.400	71	50.900	61.900
5	OK	75	415	330-6626	yes	no	0	166.700	113	28.340	148.300
6	AL	118	510	391-8027	yes	no	0	223.400	98	37.980	220.600
7	MA	121	510	355-9993	no	yes	24	218.200	88	37.090	348.500
8	MO	147	415	329-9001	yes	no	0	157.000	79	26.690	103.100
9	LA	117	408	335-4719	no	no	0	184.500	97	31.370	351.600
10	WV	141	415	330-8173	yes	yes	37	258.600	84	43.960	222.000

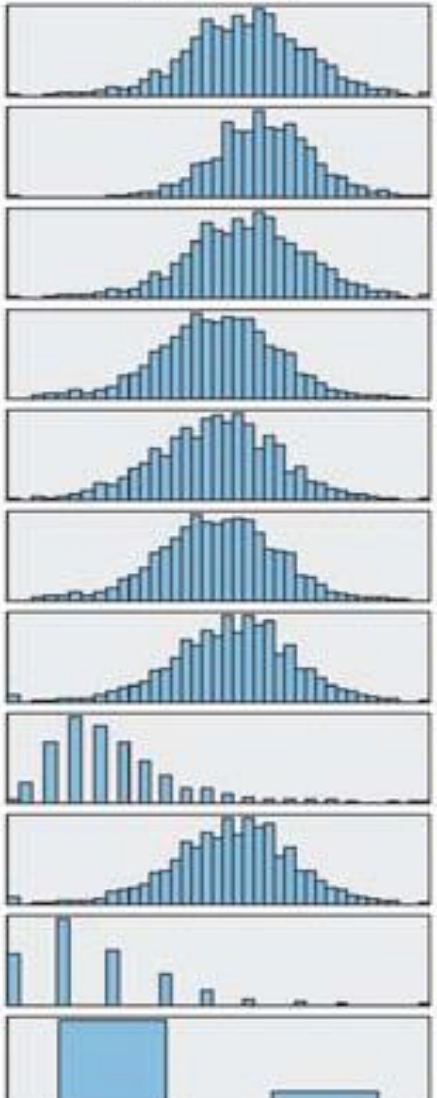
	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls	Churn
1	99	16.780	244.700	91	11.010	10.000	3	2.700	1	False
2	103	16.620	254.400	103	11.450	13.700	3	3.700	1	False
3	110	10.300	162.600	104	7.320	12.200	5	3.290	0	False
4	88	5.260	196.900	89	8.860	6.600	7	1.780	2	False
5	122	12.610	186.900	121	8.410	10.100	3	2.730	3	False
6	101	18.750	203.900	118	9.180	6.300	6	1.700	0	False
7	108	29.620	212.600	118	9.570	7.500	7	2.030	3	False
8	94	8.760	211.800	96	9.530	7.100	6	1.920	0	False
9	80	29.890	215.800	90	9.710	8.700	4	2.350	1	False
10	111	18.870	326.400	97	14.690	11.200	5	3.020	0	False

Summarization and visualization of the *churn* data set

Field	Sample Graph	Type	Min	Max	Mean	Std. Dev	Skewn...	Median	Mode	Unique	Valid
State		Set	--	--	--	--	--	--	WV	51	3333
Account Length		Range	1	243	101.065	39.822	0.097	101	105	--	3333
Area Code		Set	408	510	--	--	--	--	415	3	3333
Intl Plan		Flag	--	--	--	--	--	--	no	2	3333
VMail Plan		Flag	--	--	--	--	--	--	no	2	3333
VMail Message		Range	0	51	8.099	13.688	1.265	0	0	--	3333
Day Mins		Range	0.000	350.800	179.775	54.467	-0.029	179.400	154.000'	--	3333
Day Calls		Range	0	165	100.436	20.069	-0.112	101	102	--	3333
Day Charge		Range	0.000	59.640	30.562	9.259	-0.029	30.500	26.180'	--	3333

Summarization and visualization of the *churn* data set

Field	Sample Graph	Type	Min	Max	Mean	Std. Dev	Skewn...	Median	Mode	Unique	Valid
Eve Mins		Range	0.000	363.700	200.980	50.714	-0.024	201.400	169.900	--	3333
Eve Calls		Range	0	170	100.114	19.923	-0.056	100	105	--	3333
Eve Charge		Range	0.000	30.910	17.084	4.311	-0.024	17.120	14.250*	--	3333
Night Mins		Range	23.200	395.000	200.872	50.574	0.009	201.200	188.200*	--	3333
Night Calls		Range	33	175	100.108	19.569	0.032	100	105	--	3333
Night Charge		Range	1.040	17.770	9.039	2.276	0.009	9.050	9.450*	--	3333
Intl Mins		Range	0.000	20.000	10.237	2.792	-0.245	10.300	10.000	--	3333
Intl Calls		Range	0	20	4.479	2.461	1.321	4	3	--	3333
Intl Charge		Range	0.000	5.400	2.765	0.754	-0.245	2.780	2.700	--	3333
CustServ Calls		Range	0	9	1.563	1.315	1.091	1	1	--	3333
Churn		Flag	--	--	--	--	--	--	False	2	3333



Feel of Churn data

- The variable *Phone* uses only seven digits.
- There are two flag variables.
- Most of our variables are continuous.
- The response variable Churn is a flag variable having two values, True and False.

EXPLORING CATEGORICAL VARIABLES

- Bar graph in shows the counts and percentages of customers who churned (true) and who did not churn (false).
- Only a minority (14.49%) of our customers have left service.
- *Our task is to identify* ***patterns in the data*** *that will help to* ***reduce the proportion of churners***.

Value	Proportion	%	Count
False		85.51	2850
True		14.49	483

EXPLORING CATEGORICAL VARIABLES

Primary reasons for performing EDA is

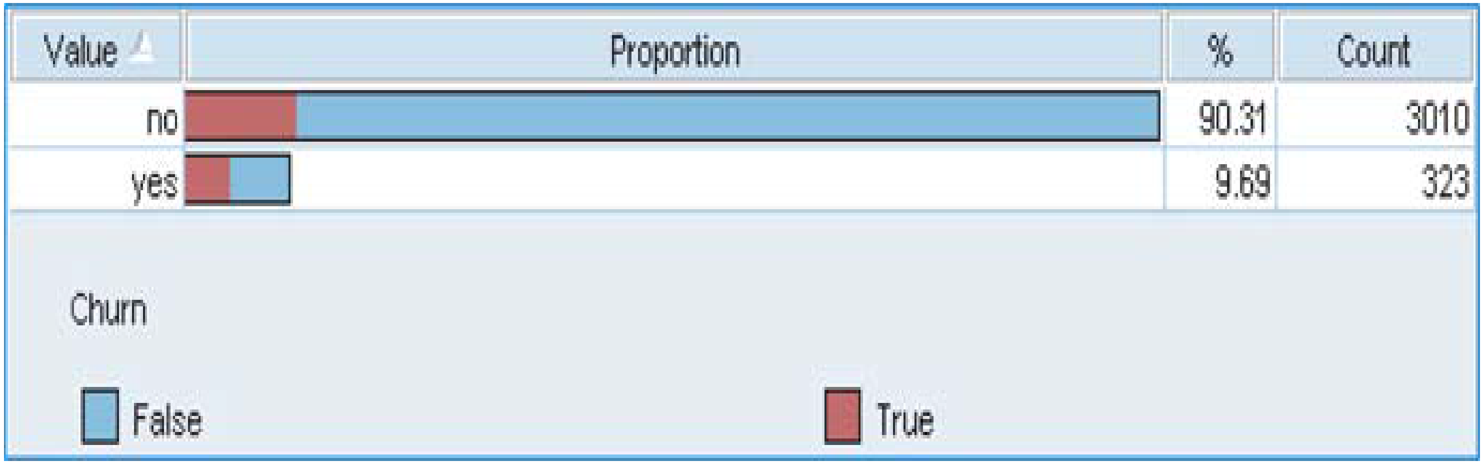
- to investigate the variables,
- examine the distributions of the categorical variables,
- look at the histograms of the numeric variables, and
- explore the relationships among sets of variables.

Overall objective to develop a model of the type of customer likely to churn

EXPLORING CATEGORICAL VARIABLES

Investigation of categorical variable *International Plan*

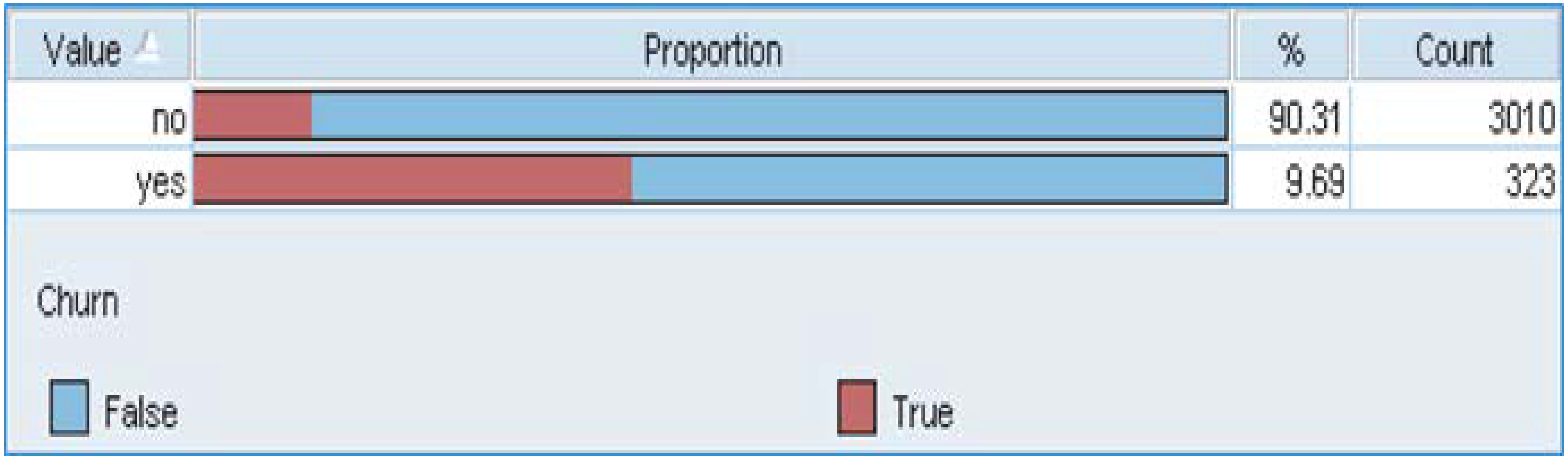
Comparison bar chart of churn proportions, by international plan participation



Greater proportion of International Plan holders are churning, but it is difficult to be sure.

EXPLORING CATEGORICAL VARIABLES

Comparison bar chart of churn proportions, by international plan participation, with equal bar length.



- ➢Clearly, those who have selected the International Plan have a greater chance of leaving the company's service

EXPLORING CATEGORICAL VARIABLES

- Graphics above tell us that International Plan holders tend to churn more frequently, but they do not **quantify the relationship**
- Use a contingency table as both variables are categorical

Contingency table of International Plan with churn

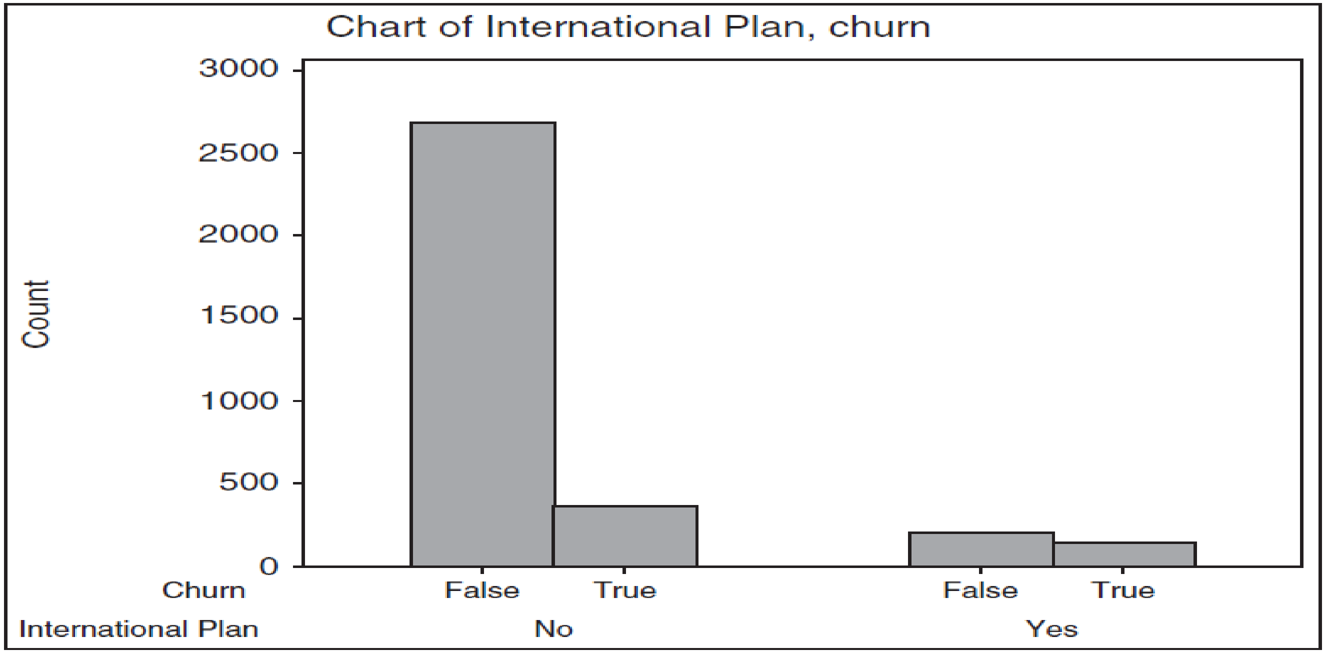
		International Plan		
		No	Yes	Total
Churn	False	2664	186	2850
	True	346	137	483
	Total	3010	323	3333

Contingency table with column percentages

		International Plan		
		No	Yes	Total
Churn	False	Count 2664 Col% 88.5%	Count 186 Col% 57.6%	Count 2850 Col% 85.5%
	True	Count 346 Col% 11.5%	Count 137 Col% 42.4%	Count 483 Col% 14.5%
	Total	3010	323	3333

EXPLORING CATEGORICAL VARIABLES

The graphical counterpart of the contingency table is the ***clustered bar chart.***

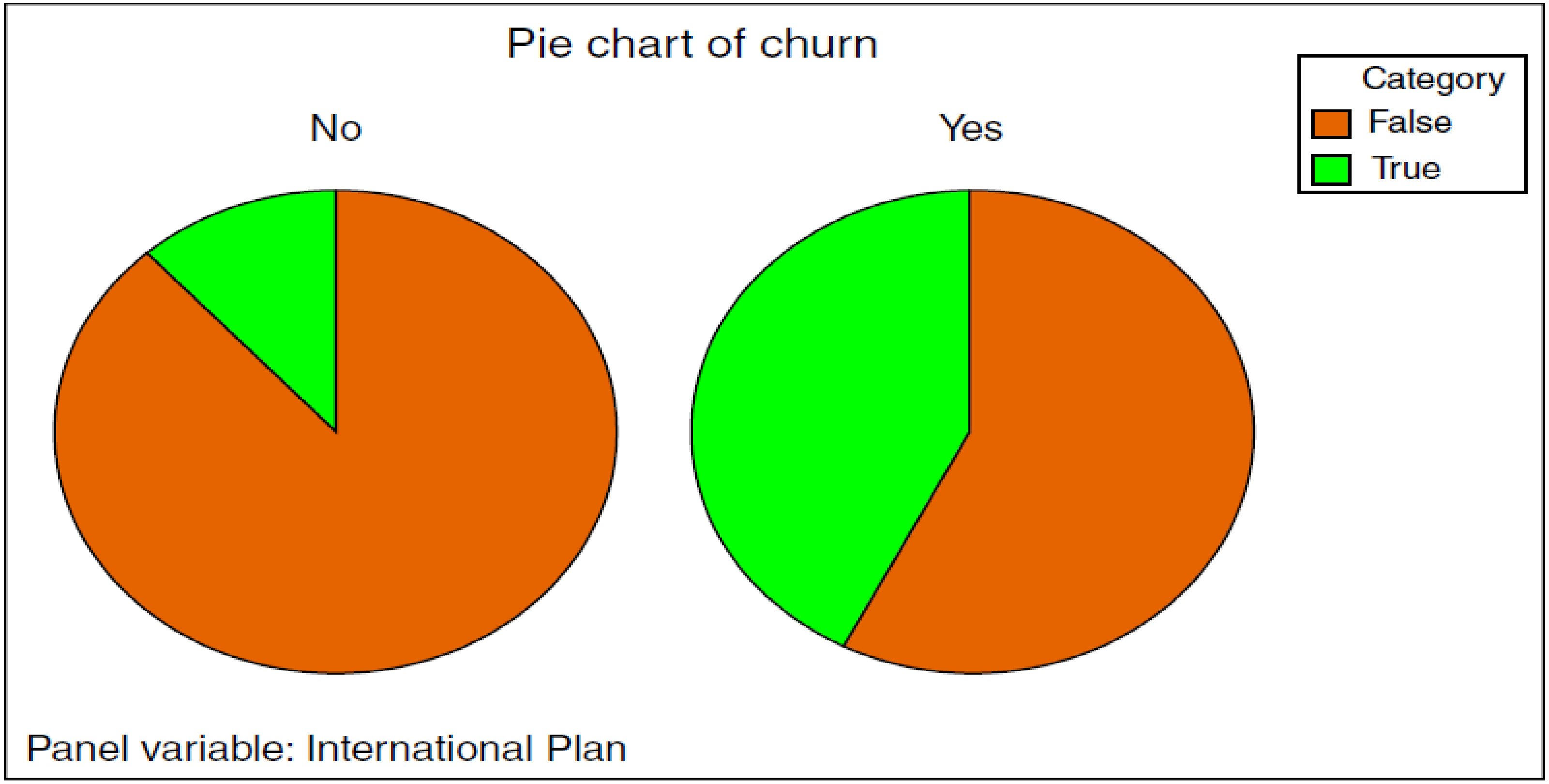


The clustered bar chart is the graphical counterpart of the contingency table.

Clearly, the proportion of churners is greater among those belonging to the **International plan**.

EXPLORING CATEGORICAL VARIABLES

Another useful graphic for comparing two categorical variables is the *comparative pie chart*.



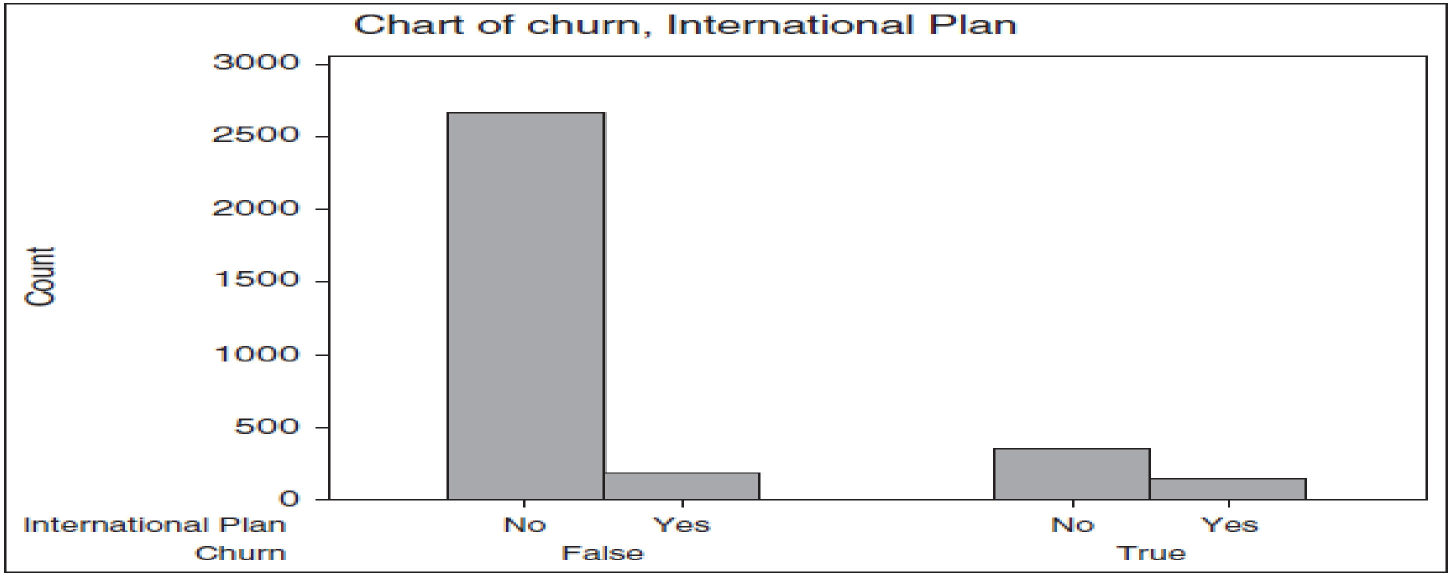
Comparative pie chart associated with Table 3.2.

EXPLORING CATEGORICAL VARIABLES

Contrast with prev. Table, the contingency table with *row percentages*

Contingency table with row percentages

		International Plan		
		No	Yes	Total
Churn	False	Count 2664 Row% 93.5%	Count 186 Row% 6.5%	2850
	True	Count 346 Row% 71.6%	Count 137 Row% 28.4%	483
	Total	Count 3010 Row% 90.3%	Count 323 Row% 9.7%	3333



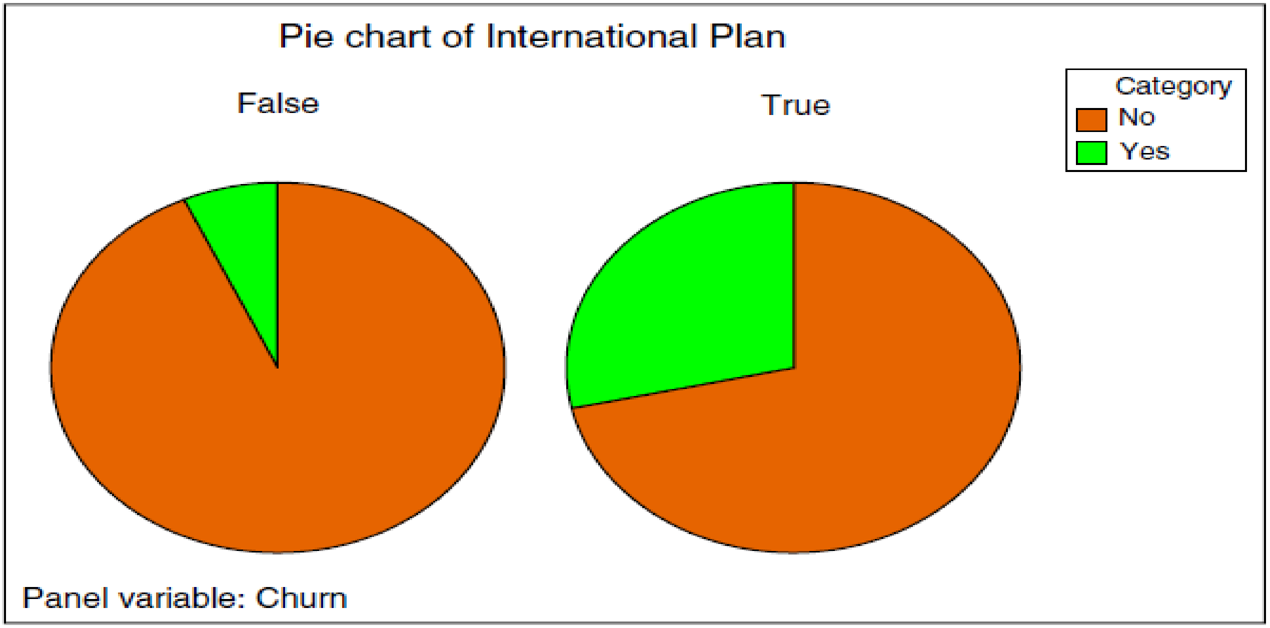
Clustered bar chart

Proportion of International Plan holders is greater among churners

EXPLORING CATEGORICAL VARIABLES

Contingency table with row percentages

		International Plan		
		No	Yes	Total
Churn	False	Count 2664 Row% 93.5%	Count 186 Row% 6.5%	2850
	True	Count 346 Row% 71.6%	Count 137 Row% 28.4%	483
	Total	Count 3010 Row% 90.3%	Count 323 Row% 9.7%	3333



Comparative pie chart

Comparative pie chart associated with above Table

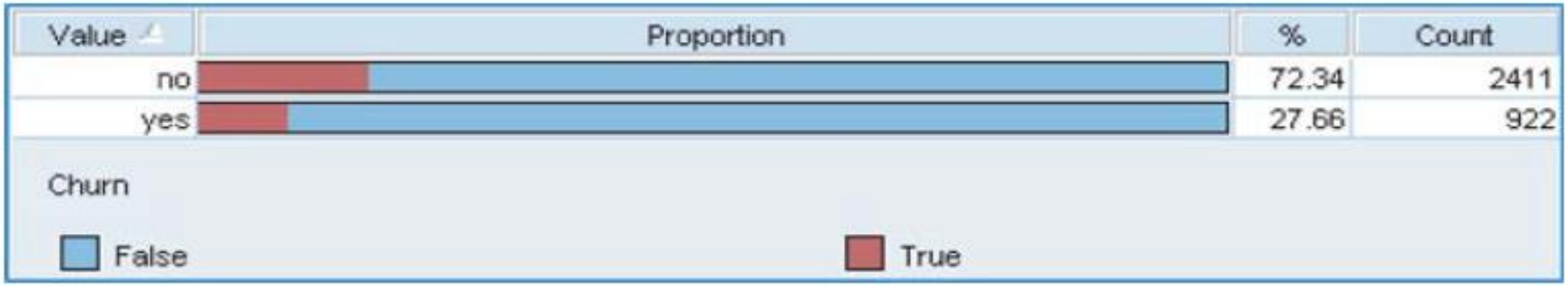
EXPLORING CATEGORICAL VARIABLES

To summarize, this EDA on the International Plan has indicated that

1. perhaps we should investigate what is it about our international plan that is inducing our customers to leave;
2. we should expect that, whatever data mining/machine learning algorithms we use to predict churn, the model will **probably include** whether or not the customer selected the International Plan.

EXPLORING CATEGORICAL VARIABLES

Let us now turn to the **Voice Mail Plan**



Those without the voice mail plan are more likely to churn.

Contingency table with column percentages for the Voice Mail Plan

		Voice Mail Plan		
		No	Yes	Total
Churn	False	Count 2008 Col% 83.3%	Count 842 Col% 91.3%	Count 2850 Col% 85.5%
	True	Count 403 Col% 16.7%	Count 80 Col% 8.7%	Count 483 Col% 14.5%
	Total	2411	922	3333

Without the Voice Mail Plan are churners, as compared to customers who do have the Voice Mail Plan.

EXPLORING CATEGORICAL VARIABLES

To summarize, this EDA on the Voice Mail Plan has indicated that

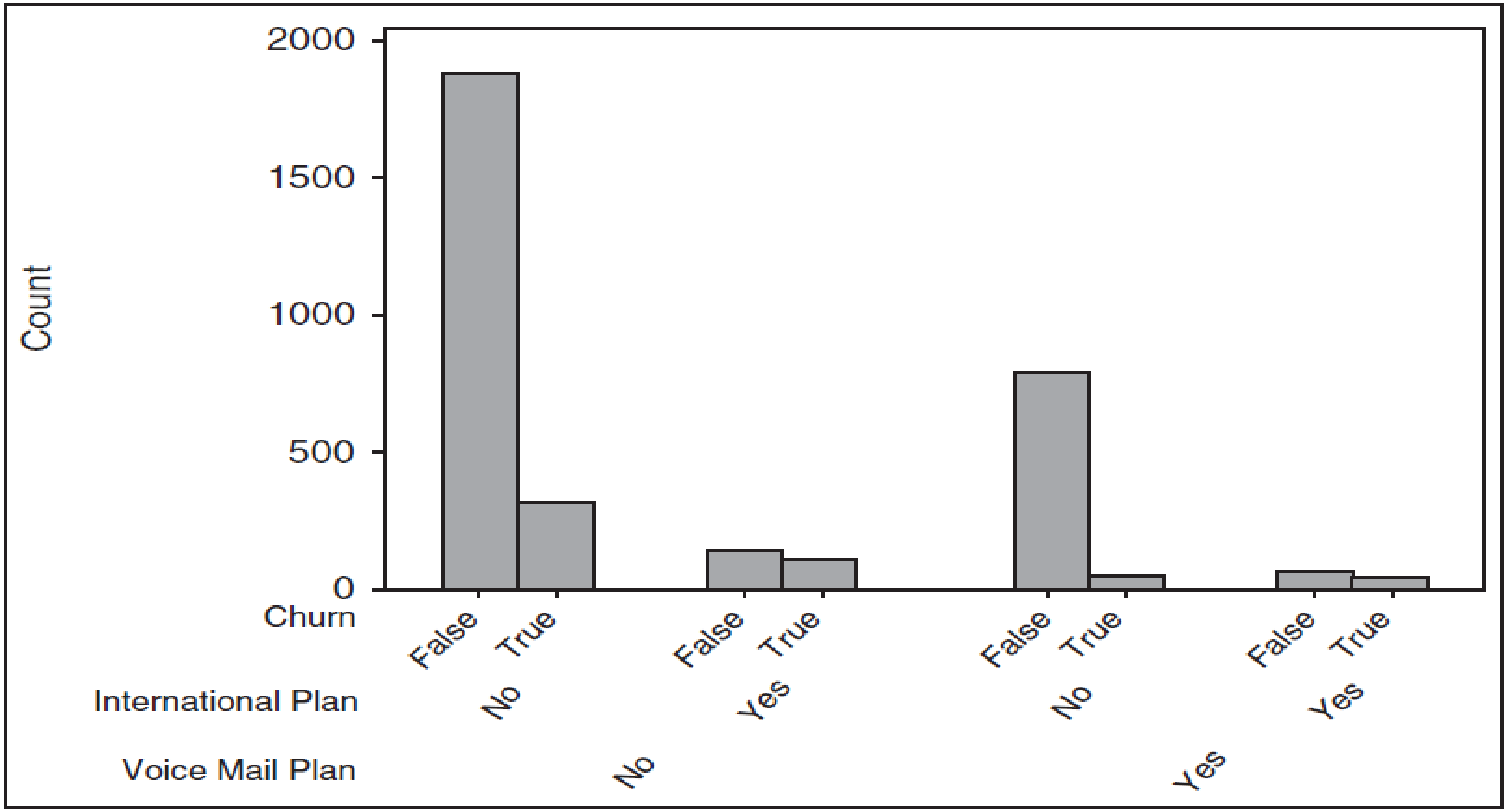
1. perhaps we should enhance our Voice Mail Plan still further, or make it easier for customers to join it, as an instrument for increasing customer loyalty;

2. whatever data mining algorithms/machine learning we use to predict churn, the model will **probably include** whether or not the customer selected the Voice Mail Plan

   - **confidence in this expectation is perhaps not quite as high as for the International Plan**

EXPLORING CATEGORICAL VARIABLES

- May also explore the ***two-way interactions*** among categorical variables with respect to ***churn***.



Multilayer clustered bar chart.

Statistics for multilayer clustered bar chart

Results for Voice Mail Plan = no

Rows: Churn   Columns: International Plan

	no	yes	All
False	1878	130	2008
True	302	101	403
All	2180	231	2411

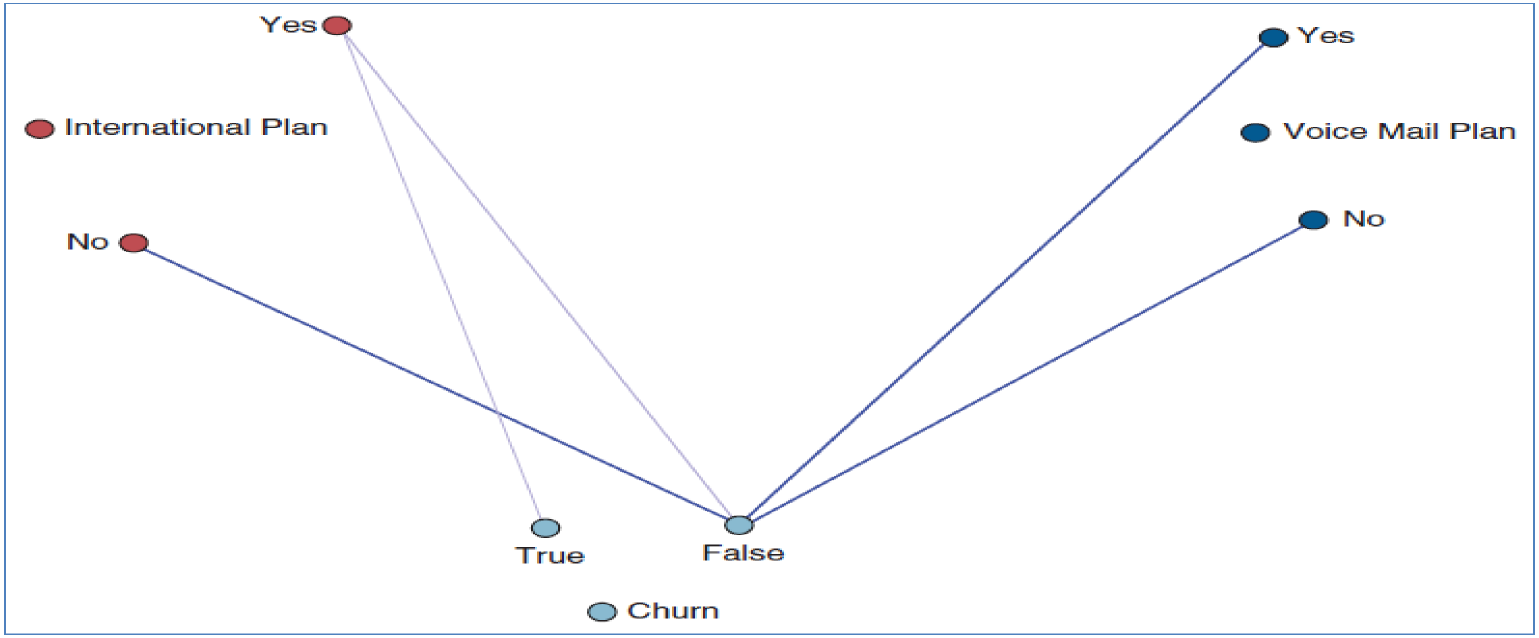
Results for Voice Mail Plan = yes

Rows: Churn   Columns: International Plan

	no	yes	All
False	786	56	842
True	44	36	80
All	830	92	922

EXPLORING CATEGORICAL VARIABLES

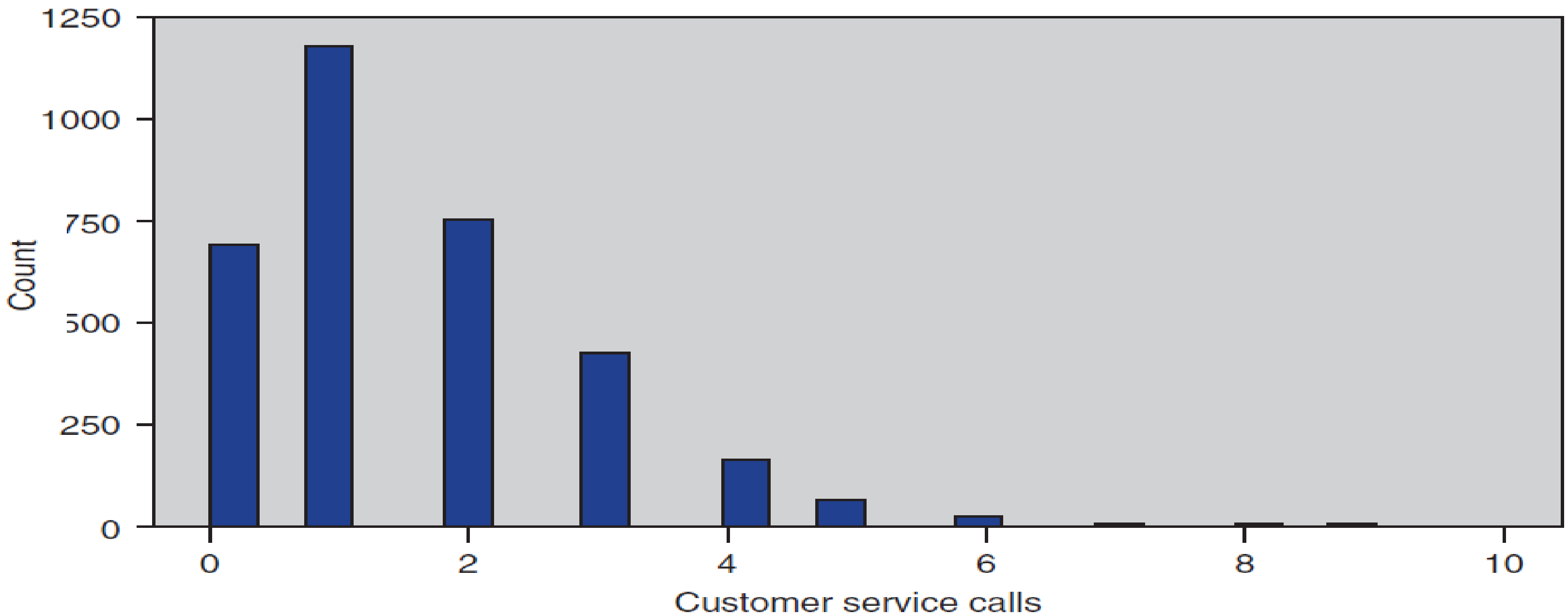
- A *directed web graph* of the relationships between International Plan holders, Voice Mail Plan holders, and churners
- Web graphs are graphical representations of the relationships between categorical variables.



Greater proportion of International Plan holders choose to churn

EXPLORING NUMERIC VARIABLES

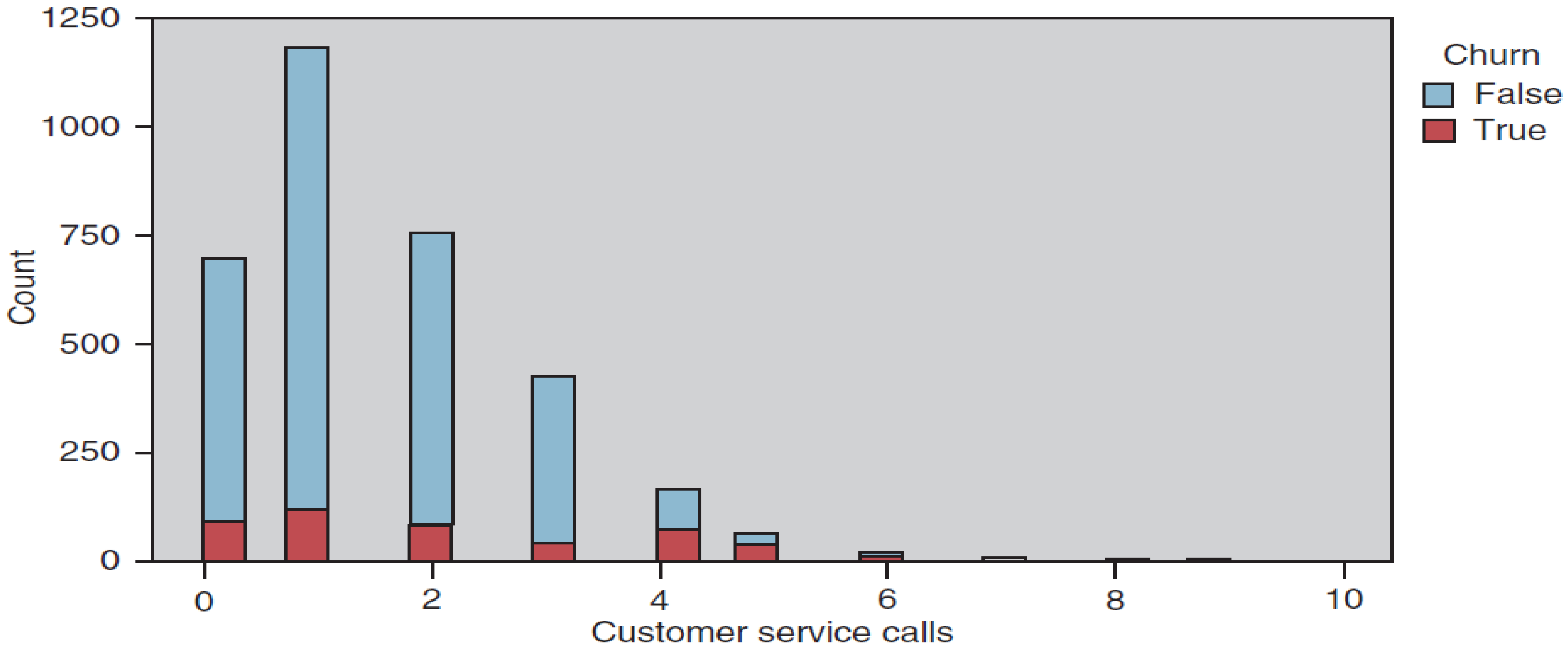
- Next, we turn to an exploration of the numeric predictive variables.
- Unfortunately, the usual type of histogram does not help us determine whether the predictor variables are associated with the target variable.



Histogram of customer service calls with no overlay.

EXPLORING NUMERIC VARIABLES

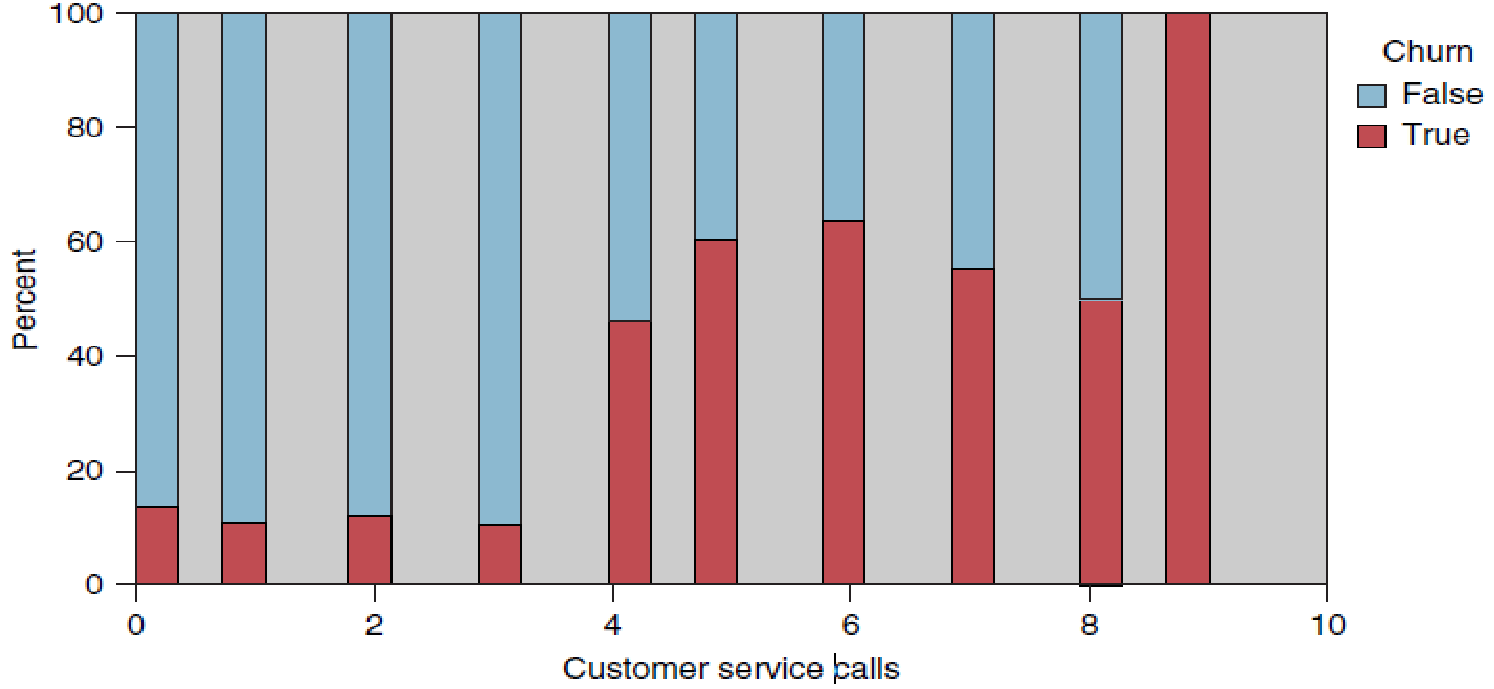
- Next, we turn to an exploration of the numeric predictive variables
- To explore whether a predictor is useful for predicting the target variable, use an overlay histogram,
- Which is a histogram where the rectangles are colored according to the values of the target variable.



Histogram of customer service calls with churn overlay.

EXPLORING NUMERIC VARIABLES

"stretching out" the rectangles that have low counts enables better definition and contrast.



"Normalized" histogram of customer service calls with churn overlay.

Customer called three times or less - lower churn rate
Customers called four or more times – higher churn rate .

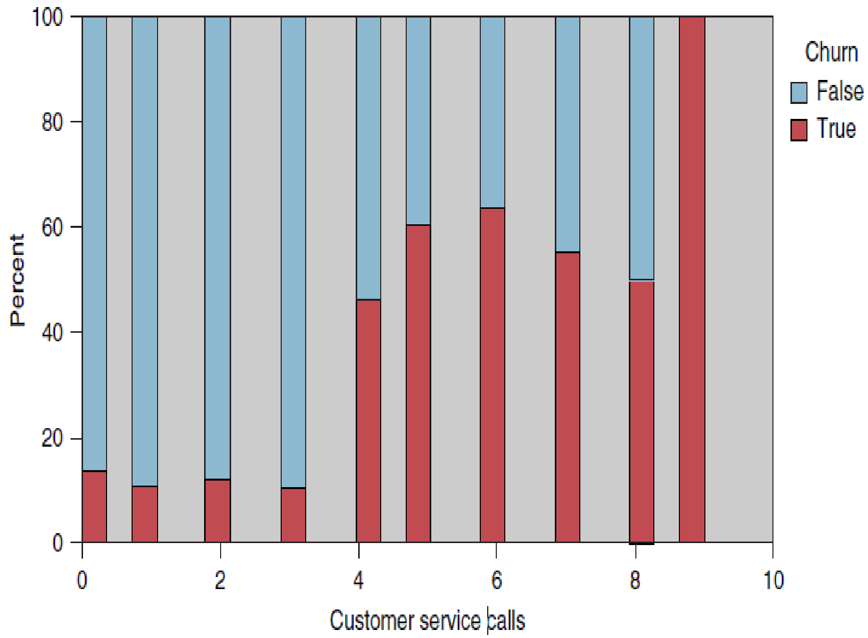
EXPLORING NUMERIC VARIABLES

This EDA on the customer service calls has indicated that

1. Carefully track the number of customer service calls made by each customer. By the third call, specialized incentives should be offered to retain customer loyalty, because, by the fourth call, the probability of churn increases greatly;
2. Whatever algorithms we use to predict churn, the model will **probably include** the number of customer service calls made by the customer.

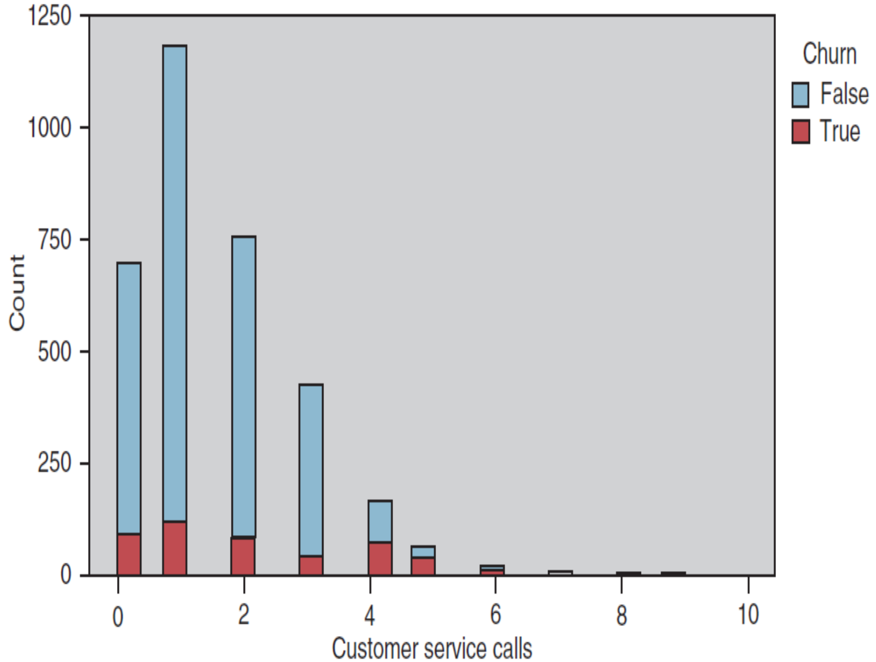
EXPLORING NUMERIC VARIABLES

Important note: Data analysts always provide a **non-normalized histogram** along with the normalized histogram, because the normalized histogram does not provide any information on the frequency distribution of the variable.



"Normalized" histogram of customer service calls with churn overlay.

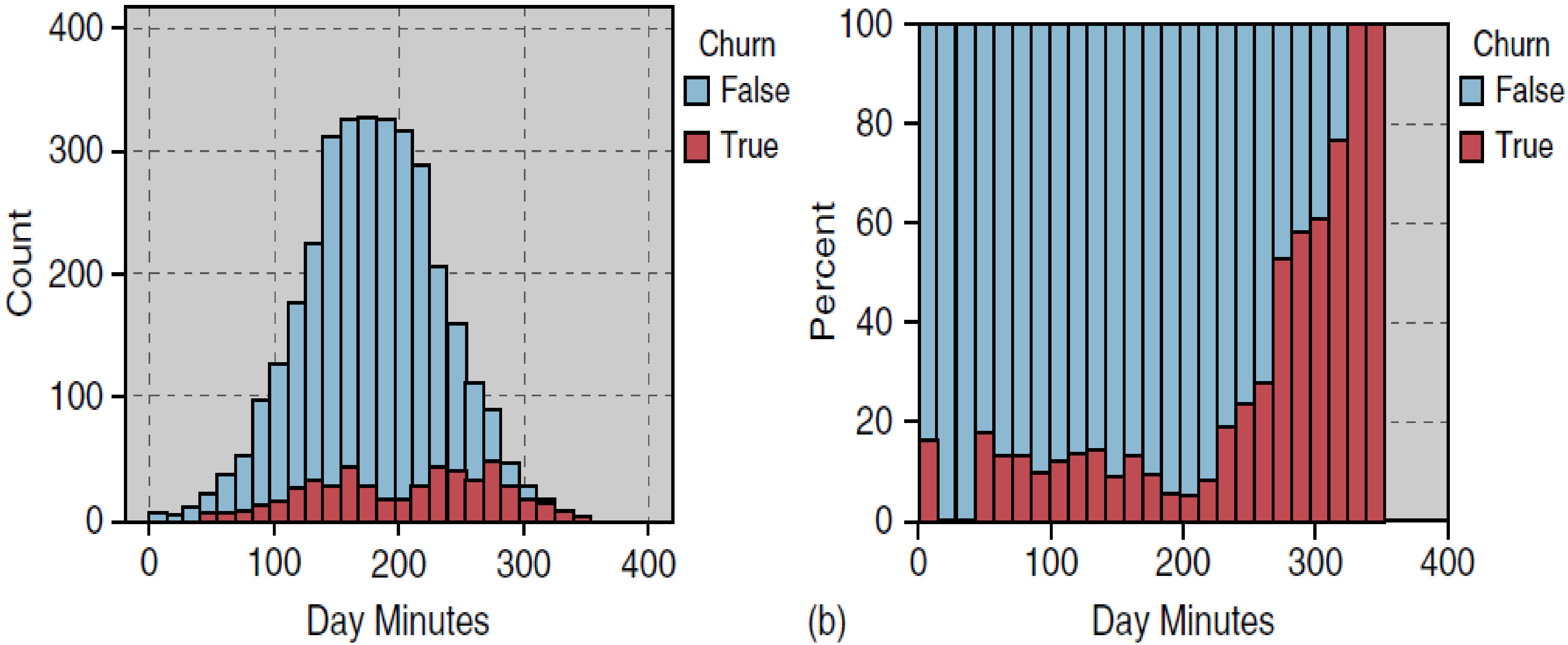
Indicates that the churn rate for customers logging nine service calls is 100%;



Histogram of customer service calls with churn overlay.

Shows that there are only two customers with this number of calls

Let us now turn to the Day Minutes



(a) Non-normalized histogram of day minutes. Normalized histogram of day

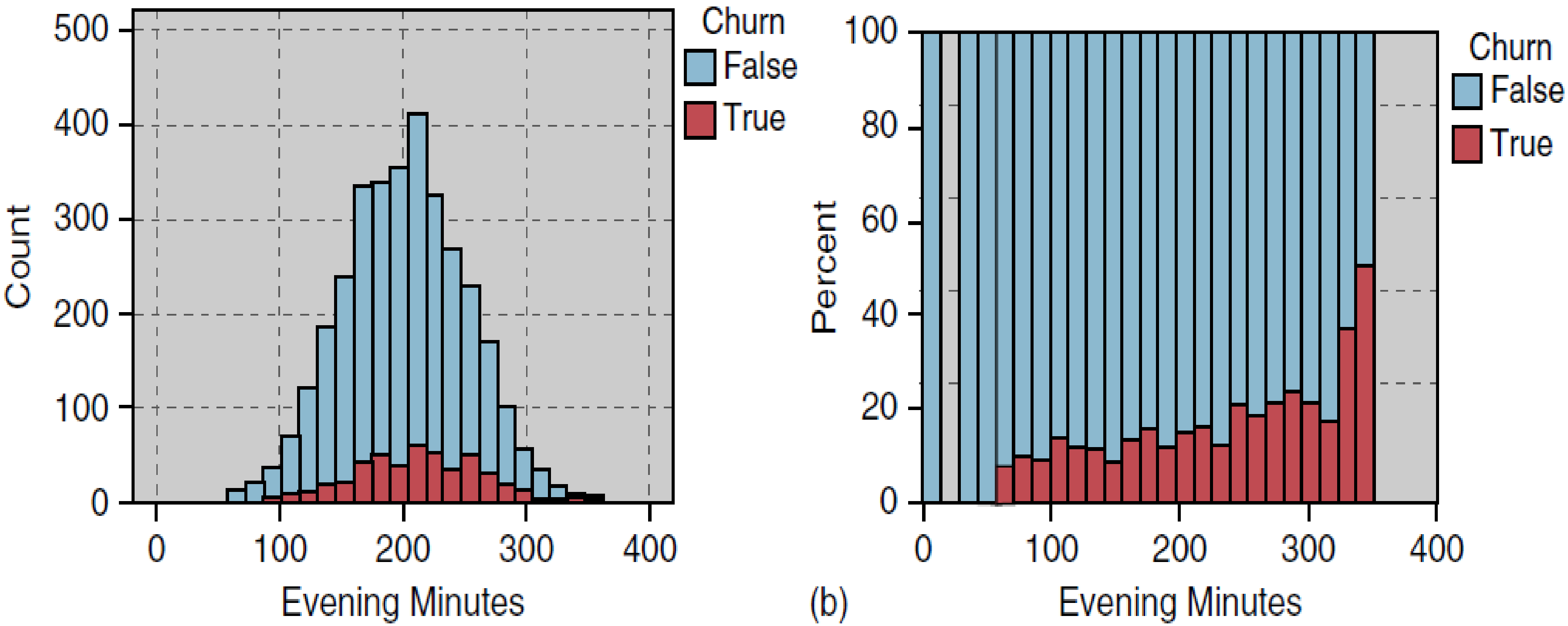
EXPLORING NUMERIC VARIABLES

The normalized histogram of ***Day Minutes*** shows that high day-users tend to churn at a higher rate. Therefore,

1. we should carefully track the number of day minutes used by each customer. As the number of day minutes passes 200, we should consider special incentives;
2. we should investigate why heavy day-users are tempted to leave;
3. we should expect that our eventual model will **include** *day minutes* as a **predictor of churn**.

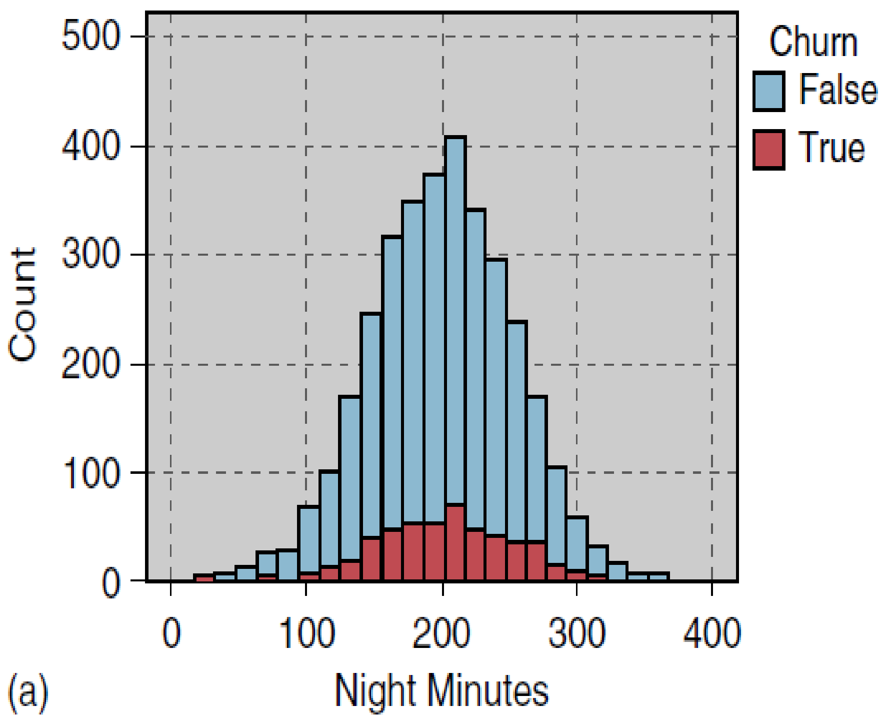
EXPLORING NUMERIC VARIABLES

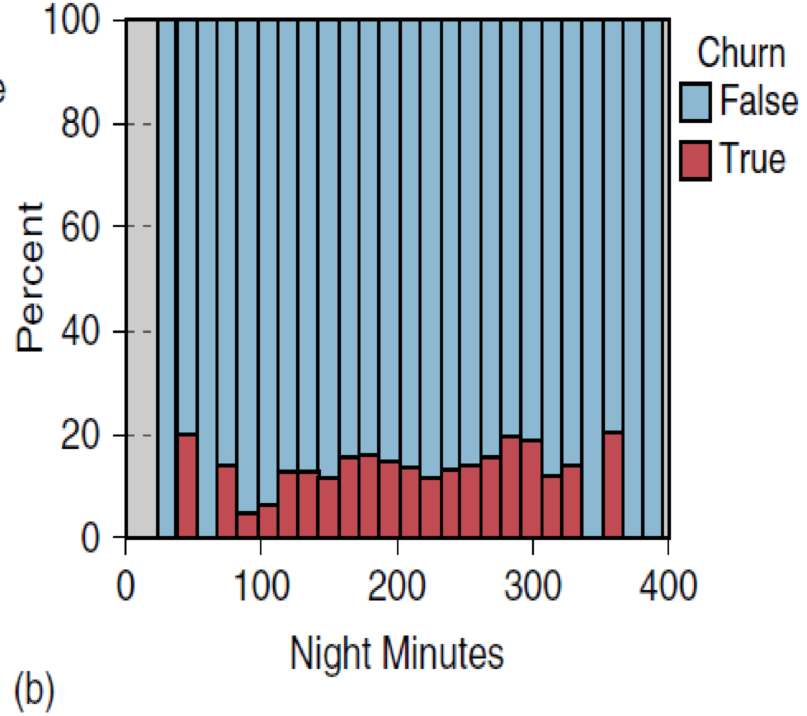
➢ **slight** tendency for customers with higher *evening minutes* to churn



(a) Non-normalized histogram of evening minutes. (b) Normalized histogram of evening minutes.

Graph indicates that there is no obvious association between churn and night minutes

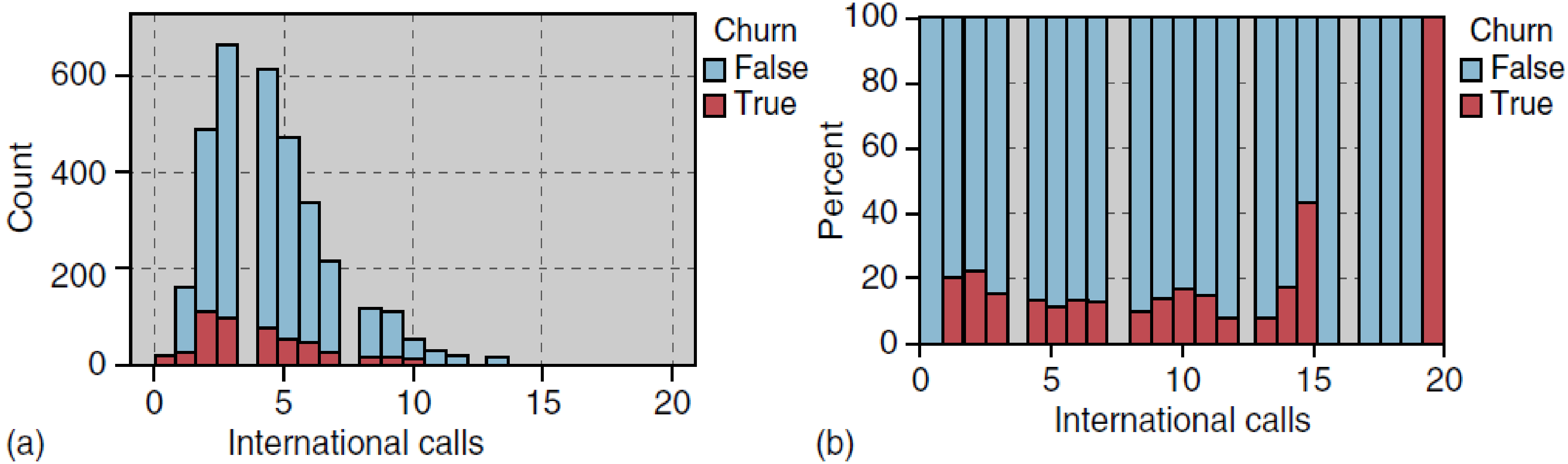




(a) Non-normalized histogram of night minutes. (b) Normalized histogram of night minutes.

EXPLORING NUMERIC VARIABLES

The lack of obvious association at the EDA stage between a predictor and a target variable is not sufficient reason to omit that predictor from the model.



(a) Non-normalized histogram of *international calls*. (b) Normalized histogram of *international calls*.

predictor International Calls with churn overlay, do not indicate strong graphical evidence of predictive importance of International Calls.

EXPLORING NUMERIC VARIABLES

- However, a *t*-test for the difference in mean number of international calls for churners and non-churners is statistically significant
- This variable is indeed useful for predicting churn:
- Churners tend to place a lower mean number of international calls

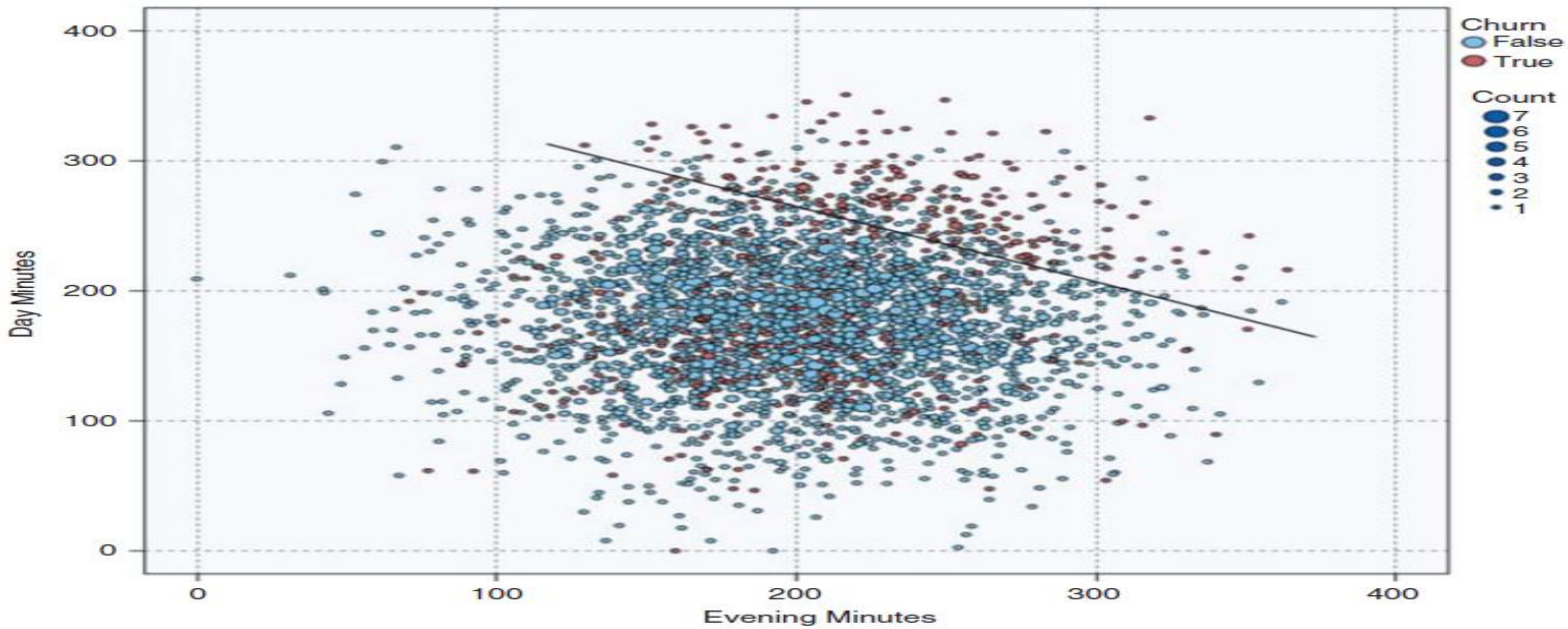
Two-Sample T-Test and CI: Intl Calls, Churn

```
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
```

- **Omitting international calls – would have committed a mistake**
- **A hypothesis test, such as this t-test lies beyond the scope of EDA**

EXPLORING MULTIVARIATE RELATIONSHIPS

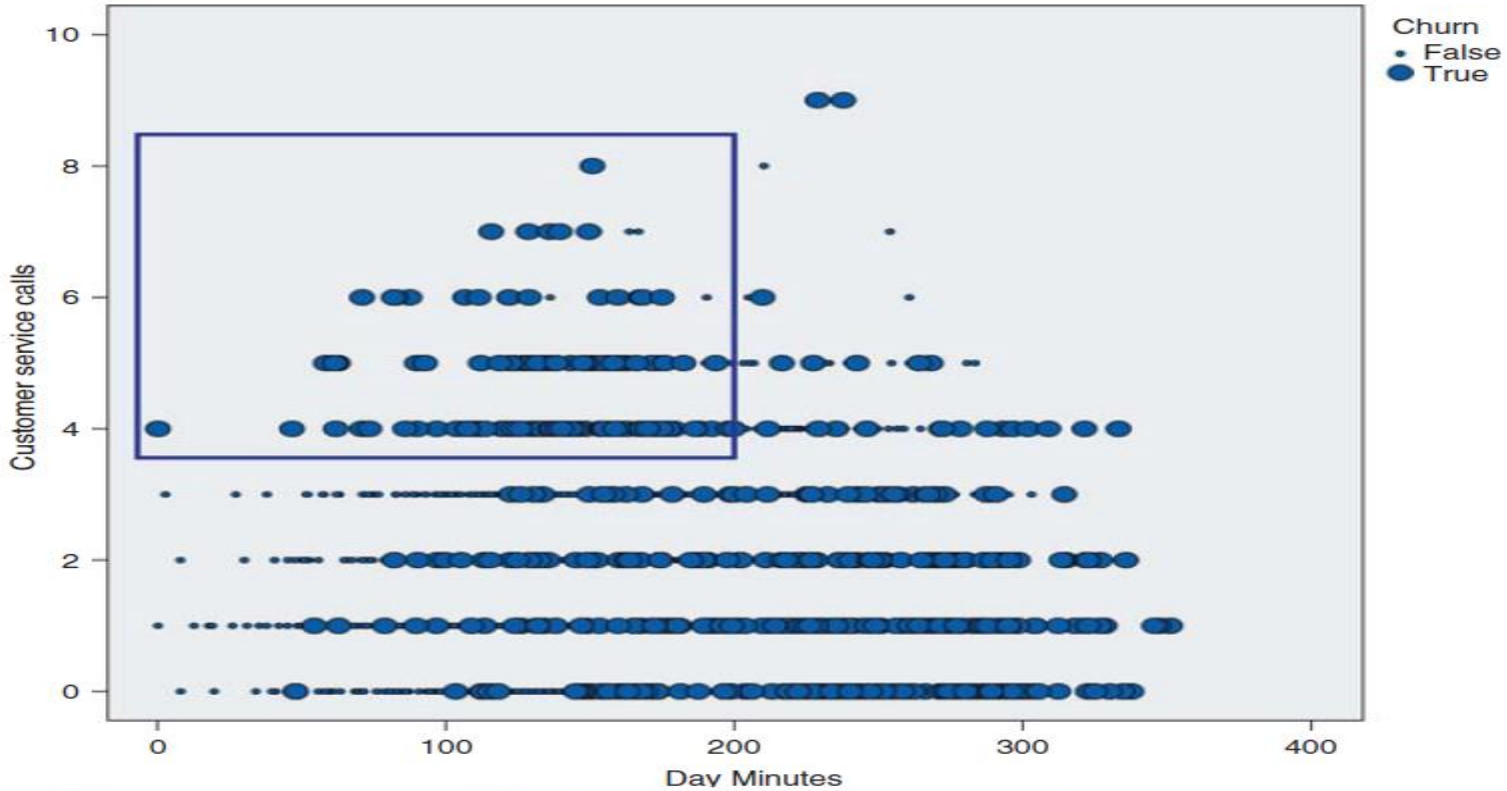
- Scatter plots can be used for examination of the possible multivariate associations



Customers with both high day minutes and high evening minutes are at greater risk of churning.

- **Records above this diagonal line** (customers high day minutes and evening minutes), - higher proportion of churners than records below line.

EXPLORING MULTIVARIATE RELATIONSHIPS



There is an interaction effect between *customer service calls* and *day minutes* with respect to churn.

- Consider the records inside the rectangle partition - indicates a high-churn area
- These records represent combination of a high number of customer service calls and a low number of day minutes used.
- This group of customers could not have been identified with **univariate exploration**

EXPLORING MULTIVARIATE RELATIONSHIPS

- Graphical EDA can uncover subsets of records that call for further investigation
- About 65% (115 of 177) of the selected records are churners

- Those with high customer service calls and low day minutes have a **65%** probability of churning

Value	Proportion	%	Count
False		35.03	62
True		64.97	115

- Compare this to the records with high customer service calls and high day minutes
- About 26% of customers with high customer service calls and high day minutes are churners

Value	Proportion	%	Count
False		74.44	67
True		25.56	23

To summarize, the strategy we implemented here is as follows:

1. Generate multivariate graphical EDA, such as scatter plots with a flag overlay.
2. Use these plots to uncover subsets of interesting records.
3. Quantify the differences by analyzing the subsets of records.

USING EDA TO UNCOVER ANOMALOUS FIELDS

- EDA can uncover strange or anomalous records or fields that the earlier data cleaning phase may have missed.
- Area code field in the contain numerals, can also be categorical variables as they can classify customers according to geographic location
- Contains only three different values for all the records, 408, 415, and 510

Value	Proportion	%	Count
408		25.14	838
415		49.65	1655
510		25.2	840

- **Would not be anomalous - customers all lived in California**

USING EDA TO UNCOVER ANOMALOUS FIELDS

- Three area codes seem to be distributed more or less evenly across all the states and the District of Columbia
- Chi-square test has a p-value of 0.608 supporting the suspicion that the area codes are distributed randomly across all the states
- Domain experts might be able to explain this type of behavior,
- Possible that the field just contains bad data
- Further communication with someone familiar with the data history, or a domain expert, is called for

	Area Code		
State	408	415	510
AK	14	24	14
AL	25	40	15
AR	13	27	15
AZ	15	36	13
CA	7	17	10
CO	25	29	12
CT	22	39	13
DC	14	27	13
DE	13	31	17
FL	12	31	20
GA	15	21	18

Cells contain: cross-tabulation of fields

Chi-square = 95.518, df = 100, probability = 0.608

BINNING BASED ON PREDICTIVE VALUE

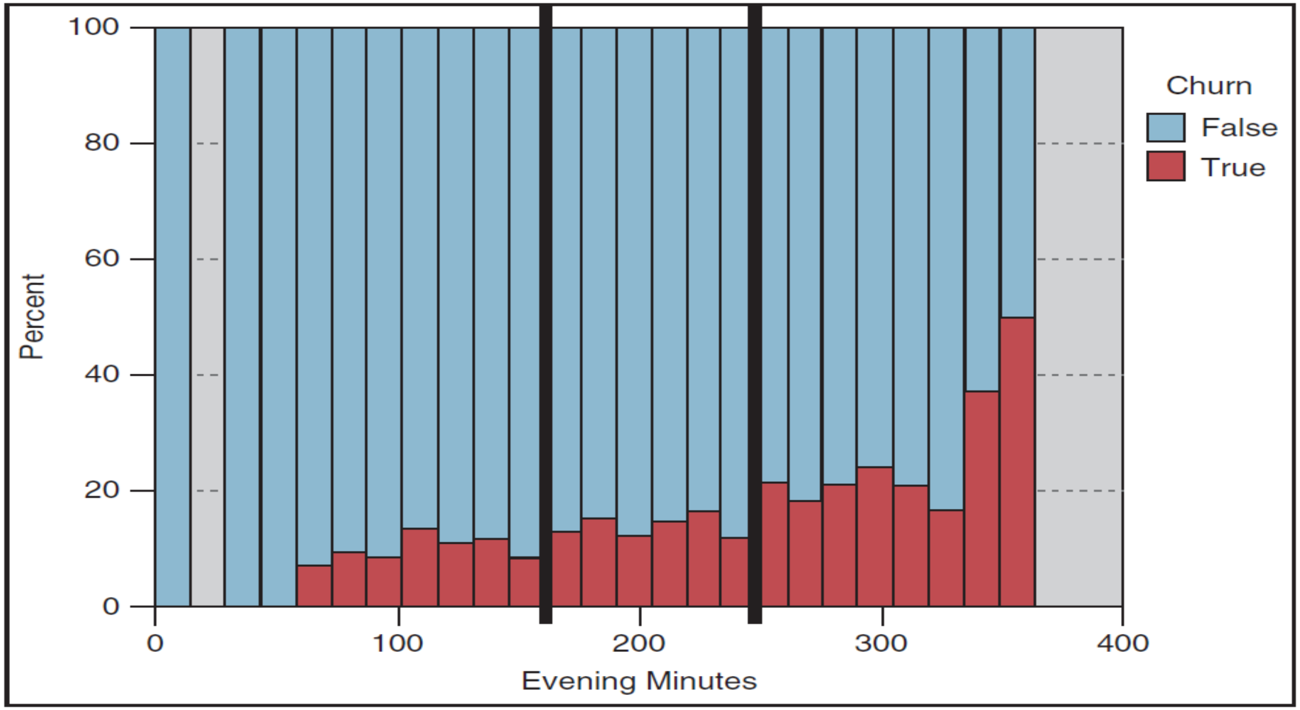
- Bin the *customer service calls* variable into two classes, *low* (fewer than four) and *high* (four or more).
- *binning of customer service calls created a flag variable with two values, high and low.*

Binning customer service calls shows difference in churn rates

		CustServPlan_Bin	
		Low	High
Churn	False	Count 2721 Col% 88.7%	Count 129 Col% 48.3%
	True	Count 345 Col% 11.3%	Count 138 Col% 51.7%

BINNING BASED ON PREDICTIVE VALUE

- trying to determine relationship between evening minutes and churn
- Can we use binning to help tease out a signal from this noise?



Binning *evening minutes* helps to tease out a signal from the noise.

BINNING BASED ON PREDICTIVE VALUE

- Binning is an art, requiring judgment.
- Where can I insert boundaries between the bins that will maximize the difference in churn proportions?
- Did the binning manage to tease out a signal?
- Can answer this by constructing a contingency table of EveningMinutes_Bin with Churn

BINNING BASED ON PREDICTIVE VALUE

Bin values for *Evening Minutes*

Bin for Categorical Variable *Evening Minutes_Bin*	Values of Numerical Variable *Evening Minutes*
Low	*Evening minutes* ≤ 160
Medium	$160 <$ *Evening minutes* ≤ 240
High	*Evening minutes* > 240

We have uncovered significant differences in churn rates among the three categorie

		EveningMinutes_Bin		
		Low	Medium	High
Churn	False	Count 618 Col% 90.0%	Count 1626 Col% 85.9%	Count 606 Col% 80.5%
	True	Count 69 Col% 10.0%	Count 138 Col% 14.1%	Count 138 Col% 19.5%

- *high* evening minutes group has nearly double the churn proportion compared to the *low* evening minutes group

DERIVING NEW VARIABLES: FLAG VARIABLES

- Deriving new variables is a data preparation activity
- EDA for usefulness of the new derived variables in predicting the target variable may be assessed

Field	Sample Graph	Type	Min	Max	Mean	Std. Dev	Skewn...	Median	Mode	Unique	Valid
State		Set	--	--	--	--	--	--	WV	51	3333
Account Length		Range	1	243	101.065	39.822	0.097	101	105	--	3333
Area Code		Set	408	510	--	--	--	--	415	3	3333
Intl Plan		Flag	--						[illegible]	2	3333
VMail Plan		Flag	--	--	--	--	--	--	no	2	3333
VMail Message		Range	0	51	8.099	13.688	1.265	0	0	--	3333
Day Mins		Range	0.000	350.800	179.775	54.467	-0.029	179.400	154.000*	--	3333
Day Calls		Range	0	165	100.436	20.069	-0.112	101	102	--	3333
Day Charge		Range	0.000	59.640	30.562	9.259	-0.029	30.500	26.180*	--	3333

Spike in distribution makes its analysis problematic.

therefore derive a flag variable

DERIVING NEW VARIABLES: FLAG VARIABLES

- Derive new *VoiceMailMessages_Flag* variables

  If *Voice Mail Messages*> 0 then

  VoiceMailMessages_Flag=1; otherwise *VoiceMailMessages_Flag* = 0.

Contingency table for *VoiceMailMessages_Flag*

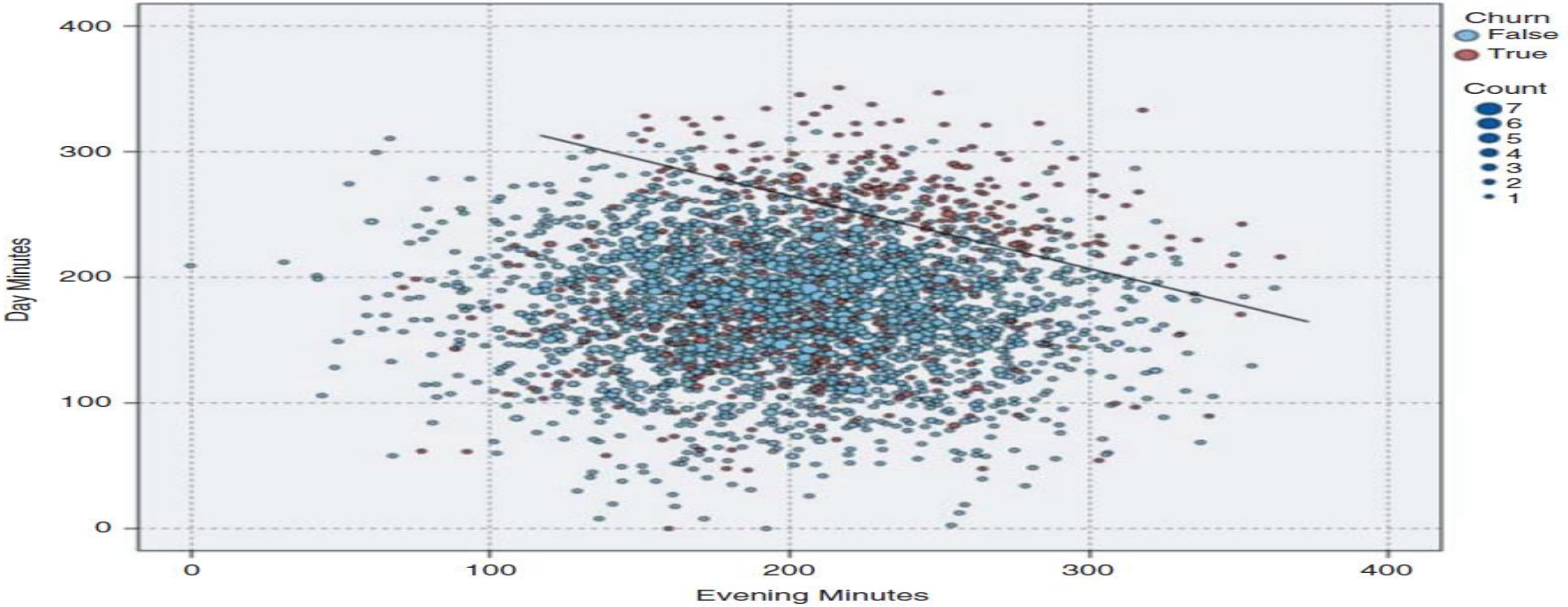
		VoiceMailMessages_Flag	
		0	1
Churn	False	Count 2008 Col% 83.3%	Count 842 Col% 91.3%
	True	Count 403 Col% 16.7%	Count 80 Col% 8.7%

Contingency table with column percentages for the Voice Mail Plan

		Voice Mail Plan		
		No	Yes	Total
Churn	False	Count 2008 Col% 83.3%	Count 842 Col% 91.3%	Count 2850 Col% 85.5%
	True	Count 403 Col% 16.7%	Count 80 Col% 8.7%	Count 483 Col% 14.5%
	Total	2411	922	3333

- Results are **exactly** the same
- *VoiceMailMessages_Flag* has **identical values** as *Voice Mail Plan*
- Derived variable is **not useful** for further analysis

DERIVING NEW VARIABLES: FLAG VARIABLES



- Both high day minutes and high evening minutes churns at a greater rate.
- Nice to quantify this claim
- Idea is to
    1. estimate the equation of the straight line;
    2. use the equation to separate the records (method portable other data set)

Estimate the equation of the line

$$\hat{y} = 400 - 0.6x$$

DERIVING NEW VARIABLES: FLAG VARIABLES

- Estimate the equation of the line

$$\hat{y} = 400 - 0.6x$$

- Create a flag variable *HighDayEveMins_Flag* as follows:

  If *Day Minutes > 400–0.6 Evening Minutes* then *HighDayEveMins_Flag* = 1; otherwise*HighDayEveMins_Flag* = 0.

- Data point above the line will have *HighDayEveMins_Flag*=1, while the data points below the line will have *HighDayEveMins_Flag*=0.

Contingency table for *HighDayEveMins_Flag*

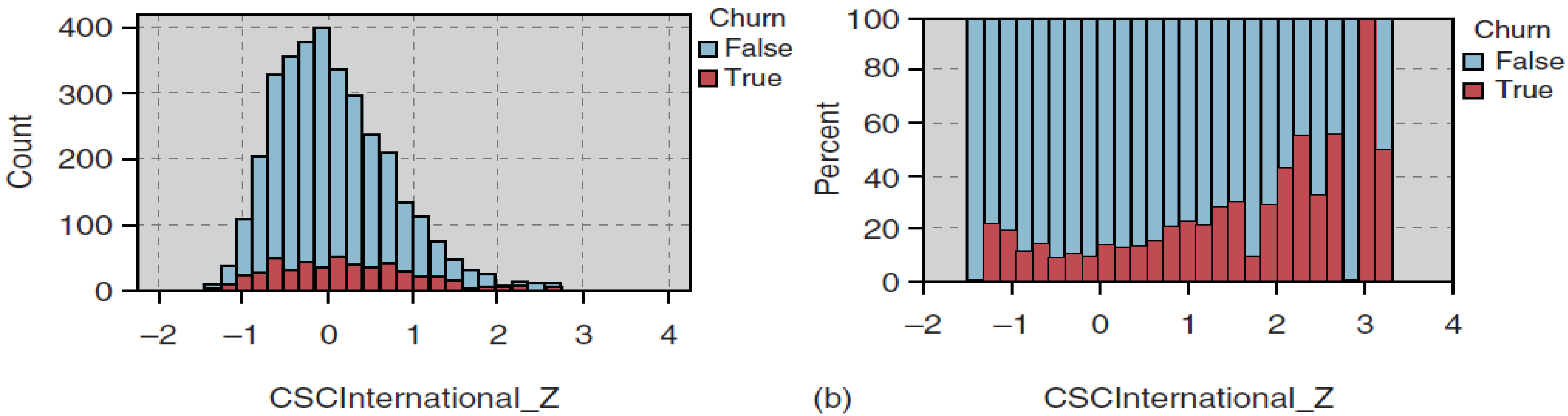
		HighDayEveMins_Flag	
		0	1
Churn	False	Count 2792 Col% 89.0%	Count 58 Col% 29.6%
	True	Count 345 Col% 11.0%	Count 138 Col% 70.4%

- Shows the highest churn proportion (70.4%)
- However, this 70.4% churn rate is restricted to a subset of fewer than 200 records

DERIVING NEW VARIABLES: NUMERICAL VARIABLES

- New numerical variable which combines Customer Service Calls and International Calls whose values will be the mean of the two fields.
- *International Calls* have a larger mean and standard deviation than *Customer Service Calls*
- *International Calls* would thereby be more heavily weighted
- We first need to standardize

$$CSCInternational\_Z = \frac{(CSC\_Z + International\_Z)}{2}$$



(a) Non-normalized histogram of *CSCInternational_Z*

(b) Normalized his togram of *CSCInternational_Z*.

- ***CSCInternational_Z* indicates that it will be useful for predicting churn.**

USING EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

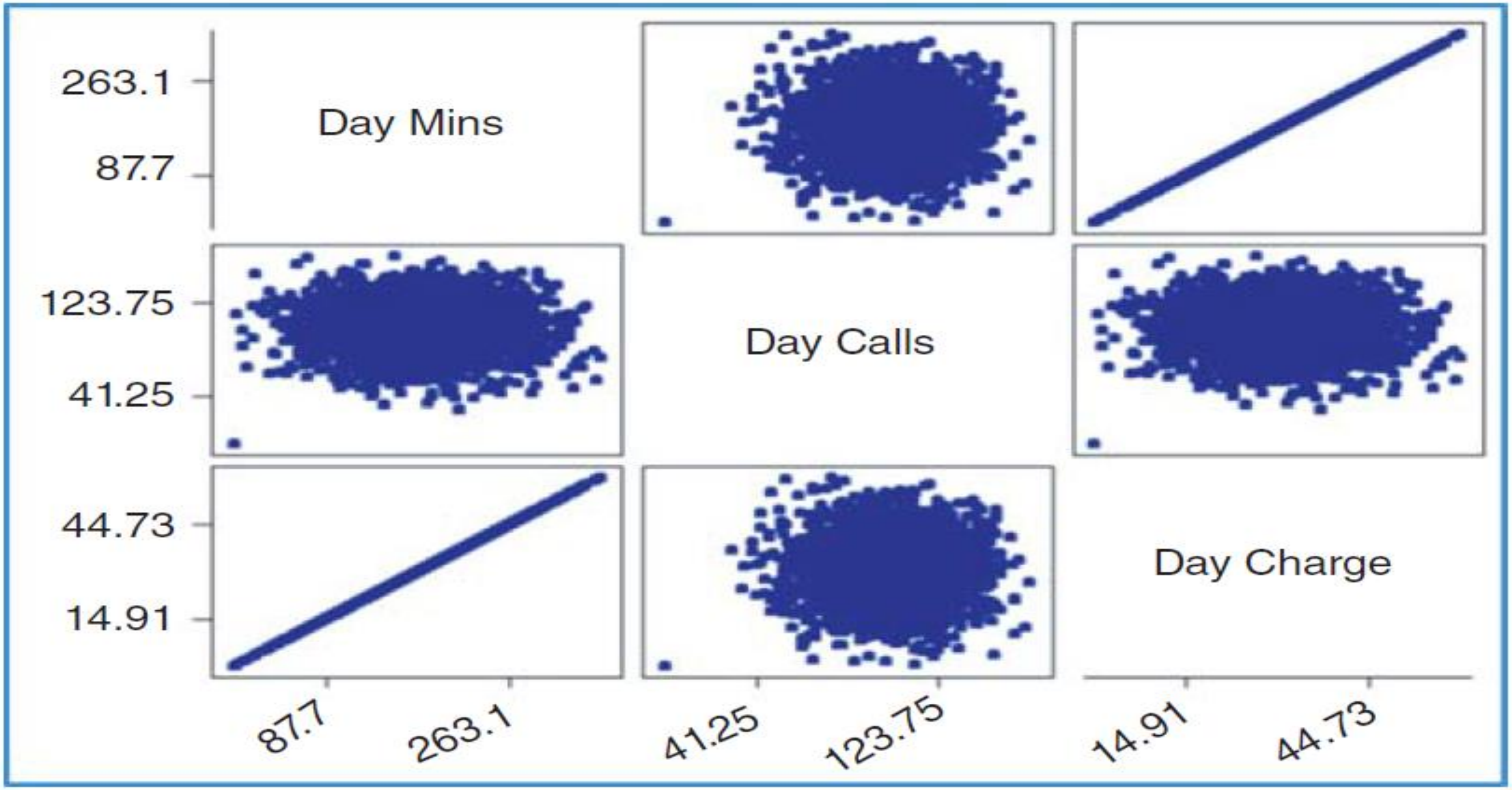
- Two variables *x* and *y* are linearly *correlated* if an increase in *x* is associated with *either an increase in y or a decrease* in *y*.
- The *correlation coefficient r* quantifies the strength and direction of the linear relationship between *x* and *y*.
- The threshold for significance of the correlation coefficient *r* depends not only on the sample size but also on data mining
- Avoid feeding correlated variables to one's data mining and statistical models.
- Using correlated variables will cause the model to become unstable and deliver unreliable results

USING EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- If two variables are correlated does not mean that we should omit one of them.
- Strategy For Handling Correlated Predictor Variables At The EDA Stage
  1. Identify any variables that are perfectly correlated (i.e., $r = 1.0$ or $r = -1.0$). Do not retain both variables in the model, but rather omit one.
  2. Identify groups of variables that are correlated with each other. Then, later, during the modeling phase, apply dimension-reduction methods, such as Principal Components Analysis (PCA) to these variables.

This strategy applies to uncovering correlation among the <u>predictors alone</u>

Correlated variables can be investigated using a *matrix plot*



Matrix plot of *day minutes*, *day calls*, and *day charge*.

USING EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

The correlation coefficient values and the *p*-values for each pairwise set of variables

Correlations and *p*-values

Correlations: Day Mins, Day Calls, Day Charge

	Day Mins	Day Calls
Day Calls	0.007	
	0.697	
Day Charge	1.000	0.007
	0.000	0.697

Cell Contents: Pearson correlation
P-Value

- No any relationship between *day minutes* and *day calls*,
- No relation between *day calls* and *day charge* – odd - expected that, as the number of calls increased, the number of minutes would tend to increase
- Linear relationship between *day minutes and day charge*

USING EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- Using Minitab's regression tool, we may express this function as the estimated regression equation: "*Day charge* equals 0.000613 plus 0.17

Minitab* regression output for *Day Charge* versus *Day Minutes

Regression Analysis: Day Charge versus Day Mins

```
The regression equation is
Day Charge =0.000613 + 0.170 Day Mins
```

Predictor	Coef	SE Coef	T	P
Constant	0.0006134	0.0001711	3.59	0.000
Day Mins	0.170000	0.000001	186644.31	0.000

S = 0.002864   R-Sq = 100.0%   R-Sq(adj) = 100.0%

- *As day charge is perfectly* correlated with *day minutes*, eliminate one of the two
- also eliminate *evening charge*, *night charge*, and *international charge*.
- proceeded to the modeling phase without first uncovering these correlations, our models may have returned incoherent results
- Reduced the number of predictors from 20 to 16
- Dimensionality of the solution space is reduced – efficiently & optimal solution

USING EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- ✓ Data analyst should turn to step 2 of the strategy, and identify any other correlated predictors, handling with principal components analysis.
- ✓ The correlation of each numerical predictor with every other numerical predictor should be checked, if feasible.
- ✓ Correlations with small *p*-values should be identified.
- ✓ Table shows - A subset of this procedure

Account length is positively correlated with *day calls*

Correlations: Account Leng, VMail Messag, Day Mins, Day Calls, CustServ Cal

	Account Length	VMail Message	Day Mins	Day Calls
VMail Message	-0.005			
	0.789			
Day Mins	0.006	0.001		
	0.720	0.964		
Day Calls	0.038	-0.010	0.007	
	0.026	0.582	0.697	
CustServ Calls	-0.004	-0.013	-0.013	-0.019
	0.827	0.444	0.439	0.274

Cell Contents: Pearson correlation
P-Value

SUMMARY OF OUR EDA

- The four *charge* fields are linear functions of the *minute* fields, and should be omitted.
- The *area code* field and/or the *state* field are anomalous, and should be omitted until further clarification is obtained.

Insights with respect to *churn* are as follows:

- Customers with the *International Plan* tend to churn more frequently.
- Customers with the *Voice Mail Plan* tend to churn less frequently.
- Customers with four or more *Customer Service Calls* churn more than four times as often as the other customers.

SUMMARY OF OUR EDA

- Customers with both high DayMinutes and high Evening Minutes tend to churn at a higher rate than the other customers.
- Customers with both high Day Minutes and high Evening Minutes churn at a rate about six times greater than the other customers.
- Customers with low Day Minutes and high Customer Service Calls churn at a higher rate than the other customers.
- Customers with lower numbers of International Calls churn at a higher rate than do customers with more international calls.
- For the remaining predictors, EDA uncovers no obvious association of churn.

Thank You !!!