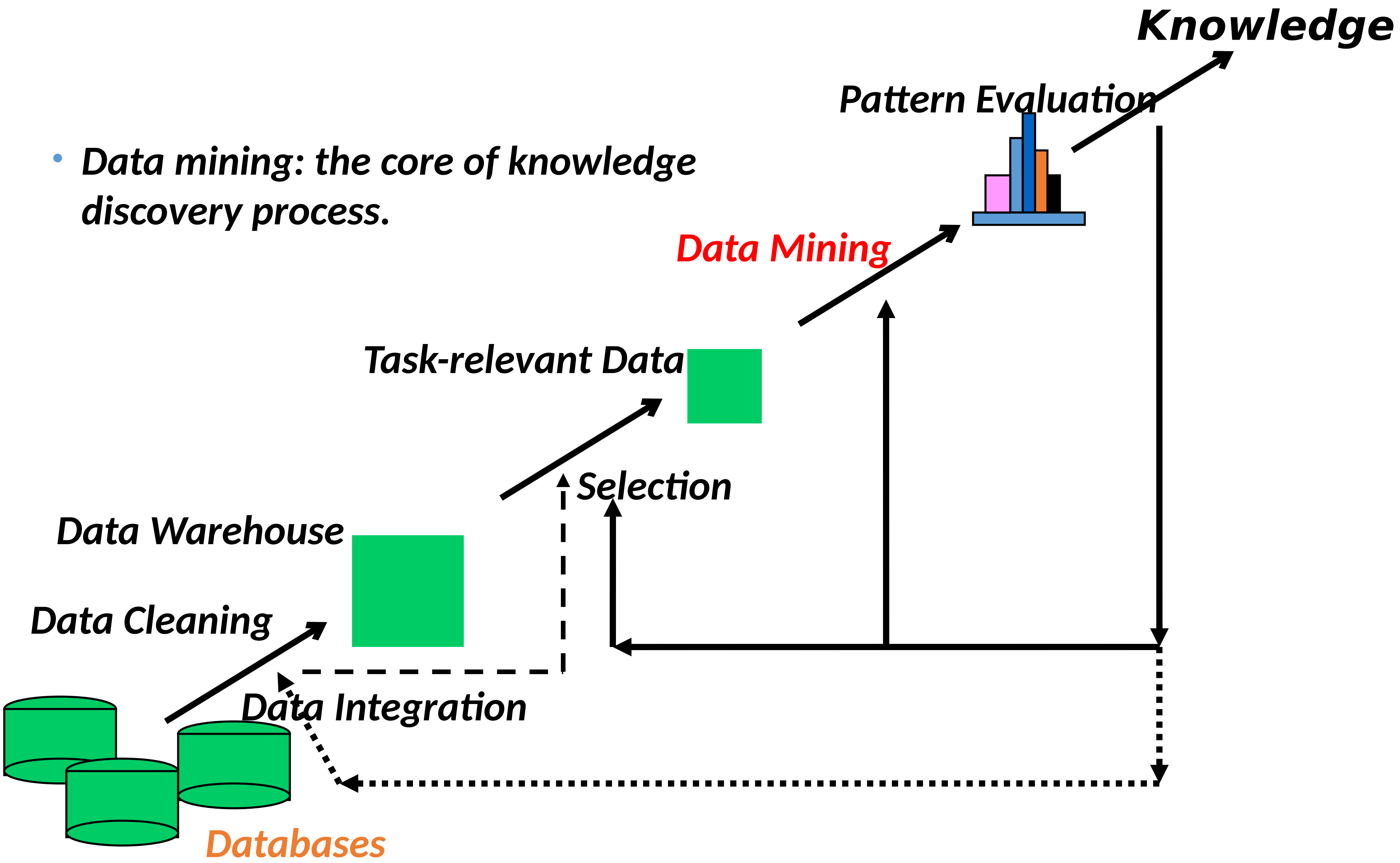
What we discussed last week?

- What is data object, attributes, types
- How to run a classification experiment?
- Data preprocessing
  - Identify outliers
  - Statistics on data
  - Data transformation
  - Skewness and normality?
  - Other data transformation methods?

KDD Process

- *Data mining: the core of knowledge discovery process.*



Exploratory Data Analysis

What is EDA?

- EDA is an approach not a set of techniques.
- EDA is a **philosophy** about how a data analysis should be carried out.

- EDA primarily uses **graphical** techniques to
  - Maximize insight into a dataset
  - Uncover underlying structure
  - Extract important variables
  - Detect outliers and anomalies
  - Test underlying assumptions
  - Determine optimal factor settings

How does EDA differ from other approaches to data analysis?

- Classical data analysis sequence
  - Problem -> Data -> Model -> Analysis -> Conclusions

- EDA data analysis sequence
  - Problem -> Data -> Analysis -> Model -> Conclusions

- Bayesian data analysis sequence
  - Problem -> Data -> Model -> Prior Distribution -> Analysis -> Conclusions

- For example, has increasing fee-structure led to decreasing market share? Hypotheses, and hypothesis testing

- How do we analyze data in the real world?

EDA vs. Classical Statistical Data Analysis

- Models
  - Classical approach imposes models on the data
  - EDA allows the data to suggest the model that best fits the data.
- Focus
  - Classical analysis focuses on the model, estimating parameters, and generating predicted values
  - EDA focuses on the data, its structure, outliers, and models suggested by the data.
- Techniques
  - Classical techniques are generally quantitative in nature (t-tests, ANOVA, chi-squared tests, and F tests.
  - EDA techniques are generally graphical (scatter plots, box plots, histograms, probability plots, etc., …)
- Rigor
  - Classical techniques are rigorous, formal and objective
  - EDA techniques are not are rigorous, are subjective, and depend on interpretation
- Treatment of the data
  - Classical techniques often map the data into a few numbers or estimates
  - EDA makes use of graphic tools and shows all of the data
- Assumptions
  - Classical techniques depend on underlying assumptions (normality)
  - EDA techniques make little or no assumptions

EDA: Getting to Know the Data Set

- Graphs, plots, and tables often uncover important relationships in data
- Example:
  - In the mobile telecommunications industry, the churn term, also known as customer attrition or subscriber churning, refers to the phenomenon of loss of a customer
- 3,333 records and 20 variables in ***churn*** data
- The two tables below shows first 10 records from churn data set
  - Simple approach looks at field values of records

	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

Attributes and Data Types

- *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*: Dichotomous categorical, yes or no,
- *Voice Mail Plan*, Dichotomous 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,
- *Total Eve Minutes*: Continuous, minutes customer used service during the evening,
- *Total Eve Calls*: Integer-valued,
- *Total Eve Charge*: Continuous, perhaps 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, perhaps based on above two variables,
- *Number of Calls to Customer Service*: Integer-valued.
- *Churn*: Target. Indicator of whether the customer has left the company (True or False).

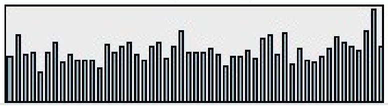
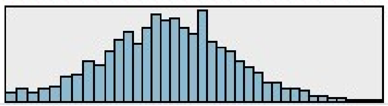
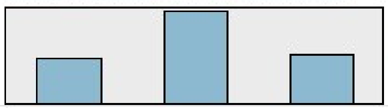
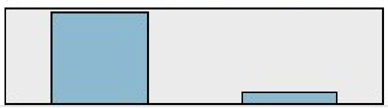
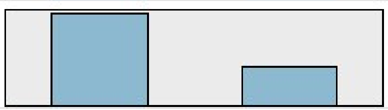
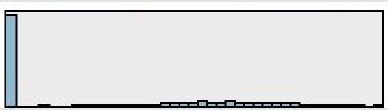
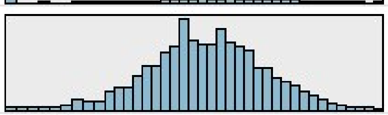
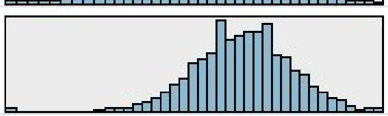
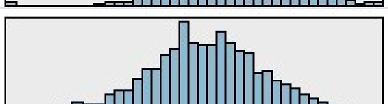
Getting to Know the Data Set *(cont'd)*

- Insights from inspecting the table:
  - The variable *Phone* uses only seven digits,
  - There are two flag variables,
  - Most of our variables are continuous, and
  - The response variable *Churn* is a flag variable having two values, *True* and *False*.
  - "**churn**" indicates customers leaving one company in favor of another company's products or services

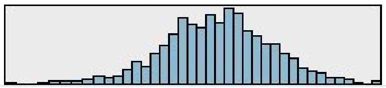
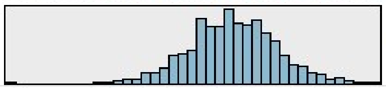
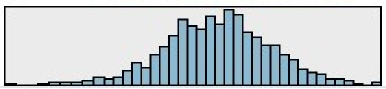
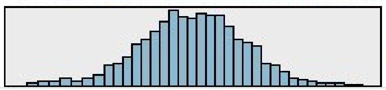
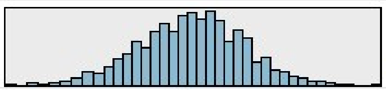
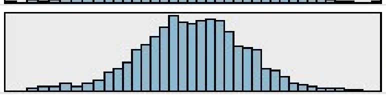
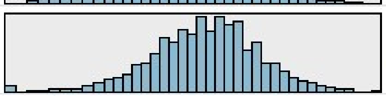
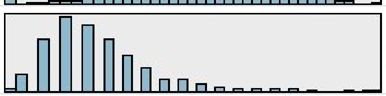
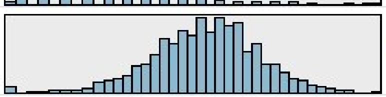
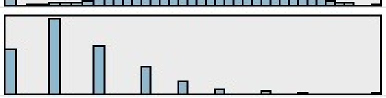
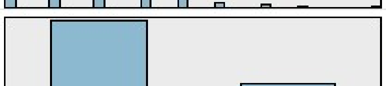
	State	Account Length	Area Code	Phone	Intl Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins
1	KS	128	415	382-4657	no	yes	25	265.100	110	45.070	197.400
2	OH	107	415	371-7191	no	yes	26	161.600	123	27.470	195.500
3	NJ	137	415	358-1921	no	no	0	243.400	114	41.380	121.200
4	OH	84	408	375-9999	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 Variables

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 Variables

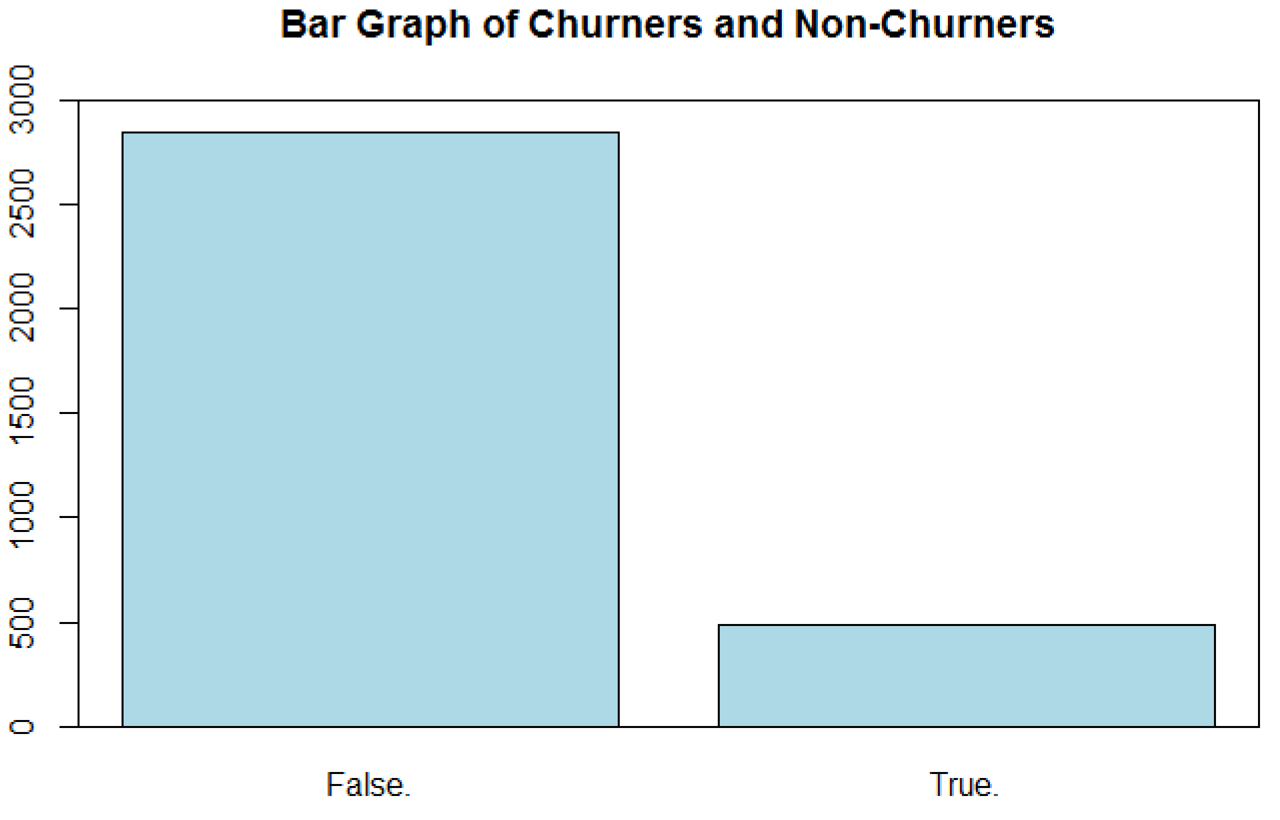
Field	Sample Graph	Type	Min	Max	Mean	Std. Dev	Skewn...	Median	Mode	Unique	Valid
Eve Mins		Range	0.000	363.700	200.980	50.714	-0.024	201.400	169.900	--	3333
Eve Calls		Range	0	170	100.114	19.923	-0.056	100	105	--	3333
Eve Charge		Range	0.000	30.910	17.084	4.311	-0.024	17.120	14.250'	--	3333
Night Mins		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

Insights

- *Vmail messages* has spike on the length
- Most quantitative variables seems normally distributed, except Intl Calls and CustServ Calls, which are right-skewed
- Unique (# of distinct field values) shows 51 for *State*, but only 3 for *Area Code* – how can this be?
- Mode for *State* is West Virginia
- International plan and voice mail plan look very similar to churn

Exploring Categorical Variables

- Bar Charts
- How many customers churned?



Exploring Categorical Variables

- General Exploratory Data Analysis Goals
  - Investigate variables
    - Examine Distributions of Categorical variables
    - Look at Histograms of numerical variables
    - Explore relationships among sets of variables
- Specific goal for ***Churn*** data mining example （ whole objective ）
  - Develop a model for the type of customer likely to churn
- Today’s software packages allow us to
  - Become familiar with the variables <u>and at the same time</u>, **begin to see which variables are associated with churn**
- Objective: Explore the data while keeping an eye on the overall
  - Bar graph below shows counts and percentages of customers who churned and did not churn

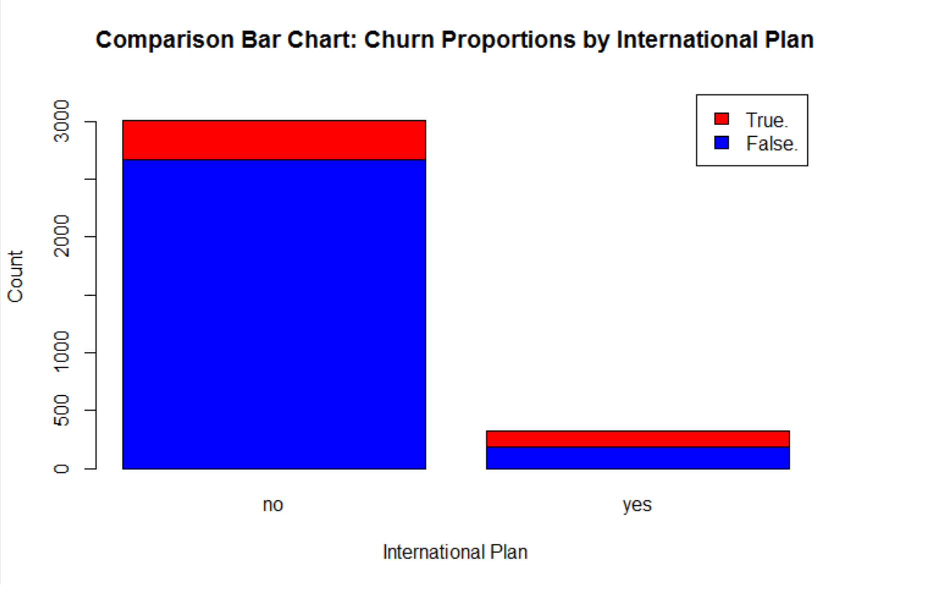
Value	Proportion	%	Count
False		85.51	2850
True		14.49	483

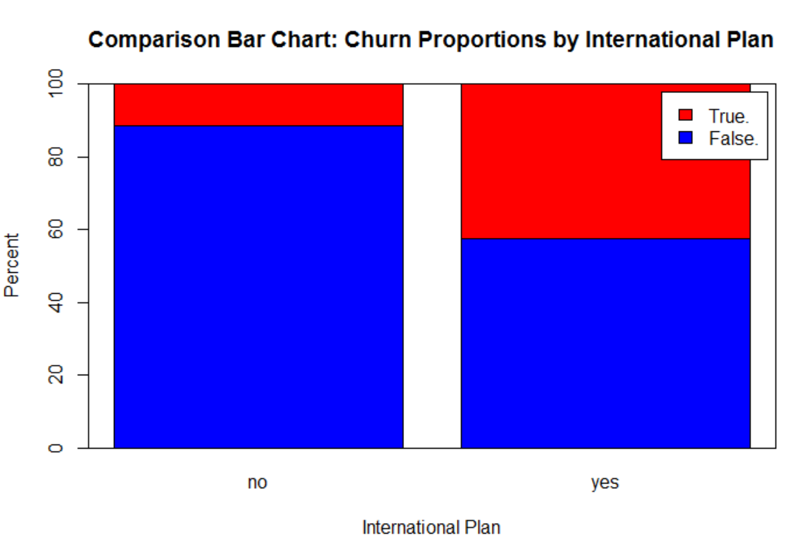
Comparing Two Categorical Variables

- How many customers churned and had international plans?
- Contingency/Crosstabulation tables and related bar charts

	International Plan	
Churn	No	Yes
False	2664	186
True	346	137

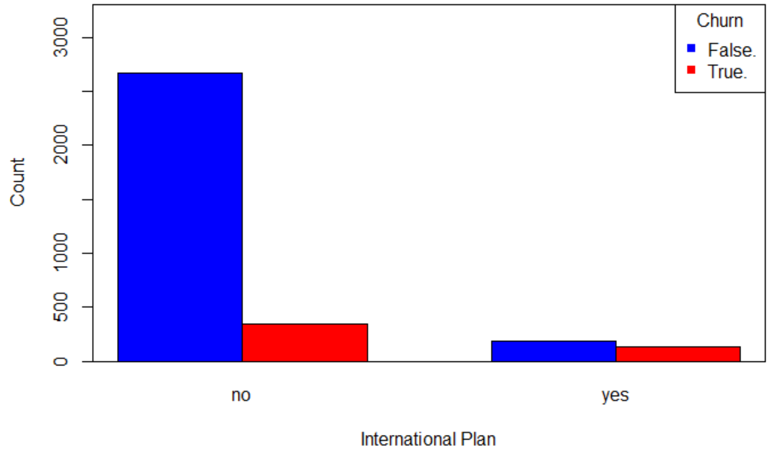
	International Plan	
Churn	No	Yes
False	88.50%	57.59%
True	11.50%	42.41%

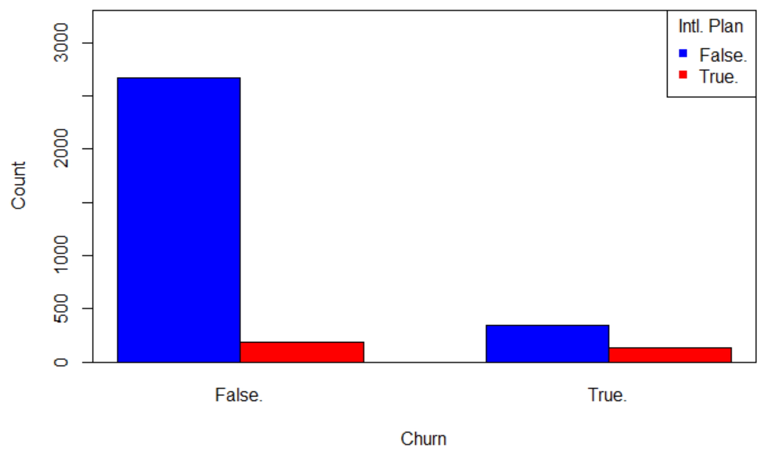




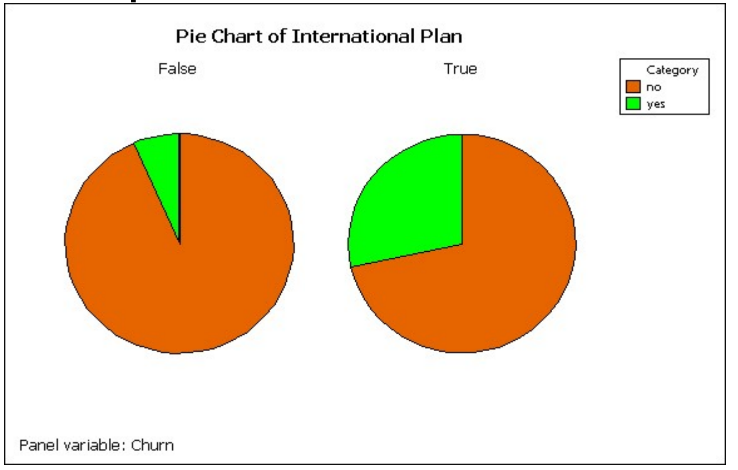
Comparing Two Categorical Variables (Other Methods)

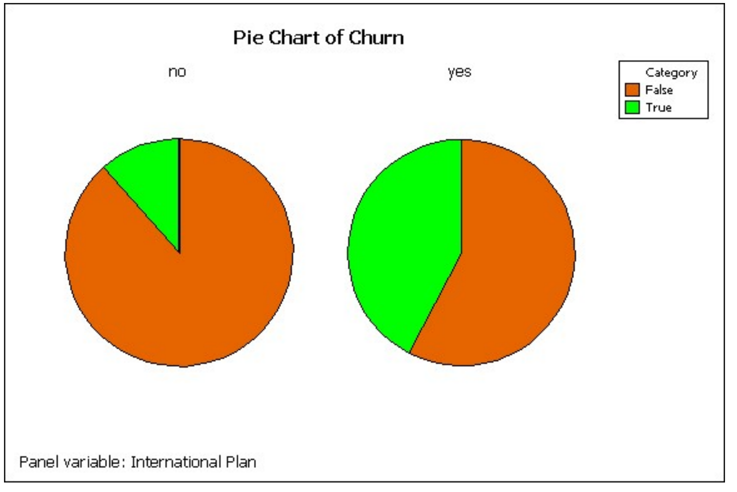
- Clustered Bar Charts





- Comparative Pie Charts



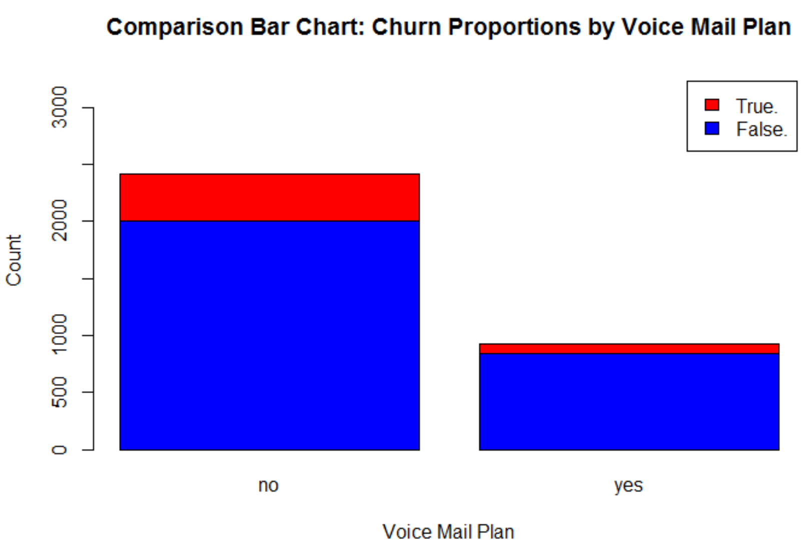


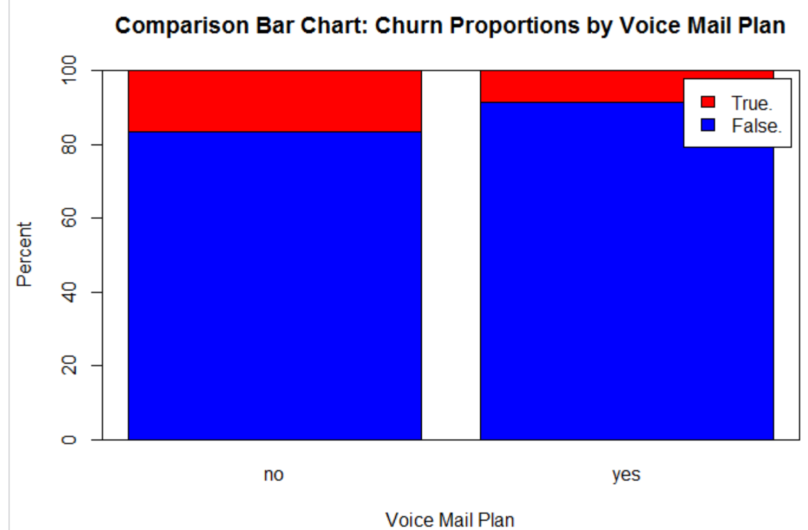
Comparing Two Categorical Variables

- How many customers churned and had **voicemail**?
- Contingency/Crosstab tables and related bar charts

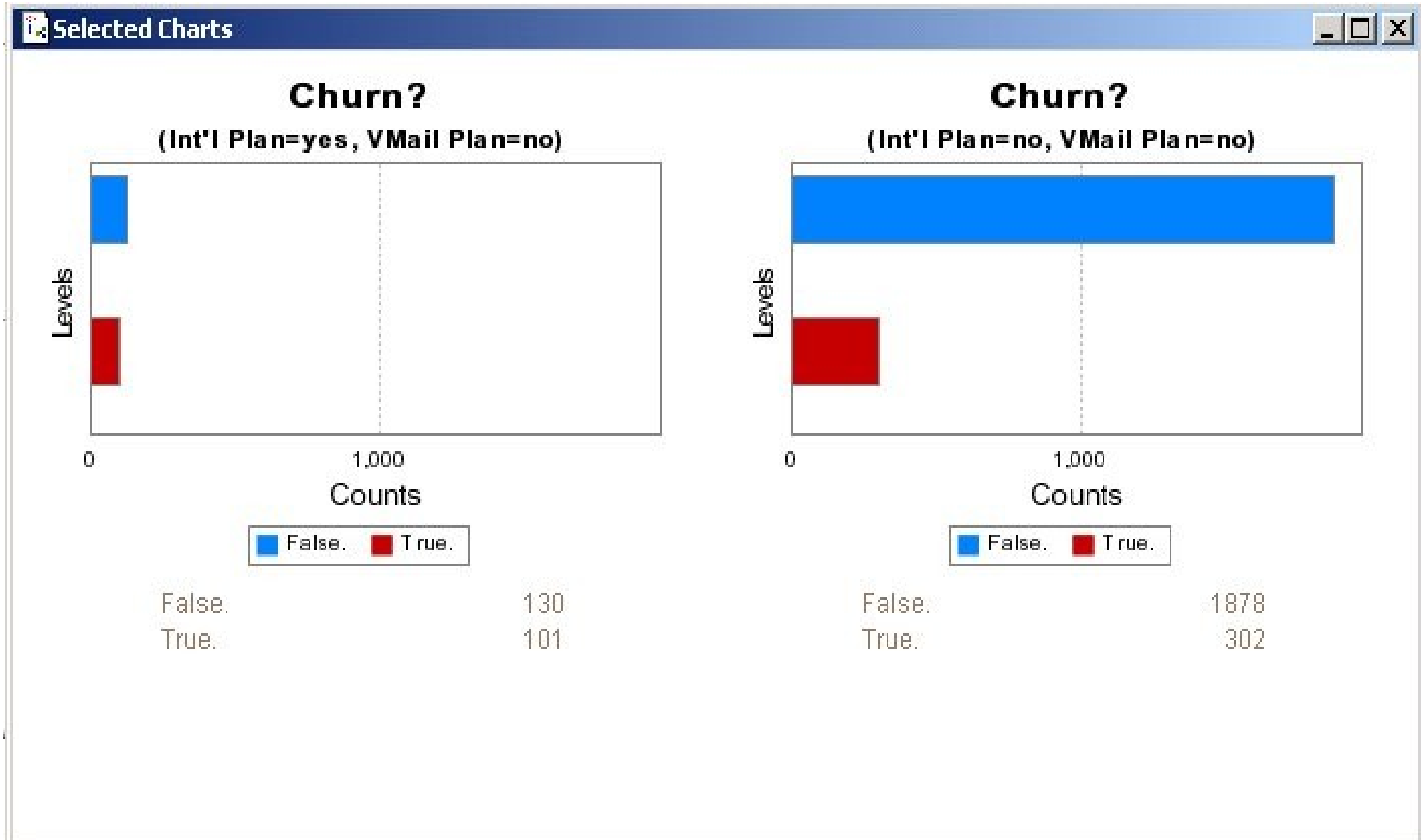
	Voice Mail Plan	
Churn	No	Yes
False	2008	842
True	403	80

	Voice Mail Plan	
Churn	No	Yes
False	83.28%	91.32%
True	16.72%	8.68%

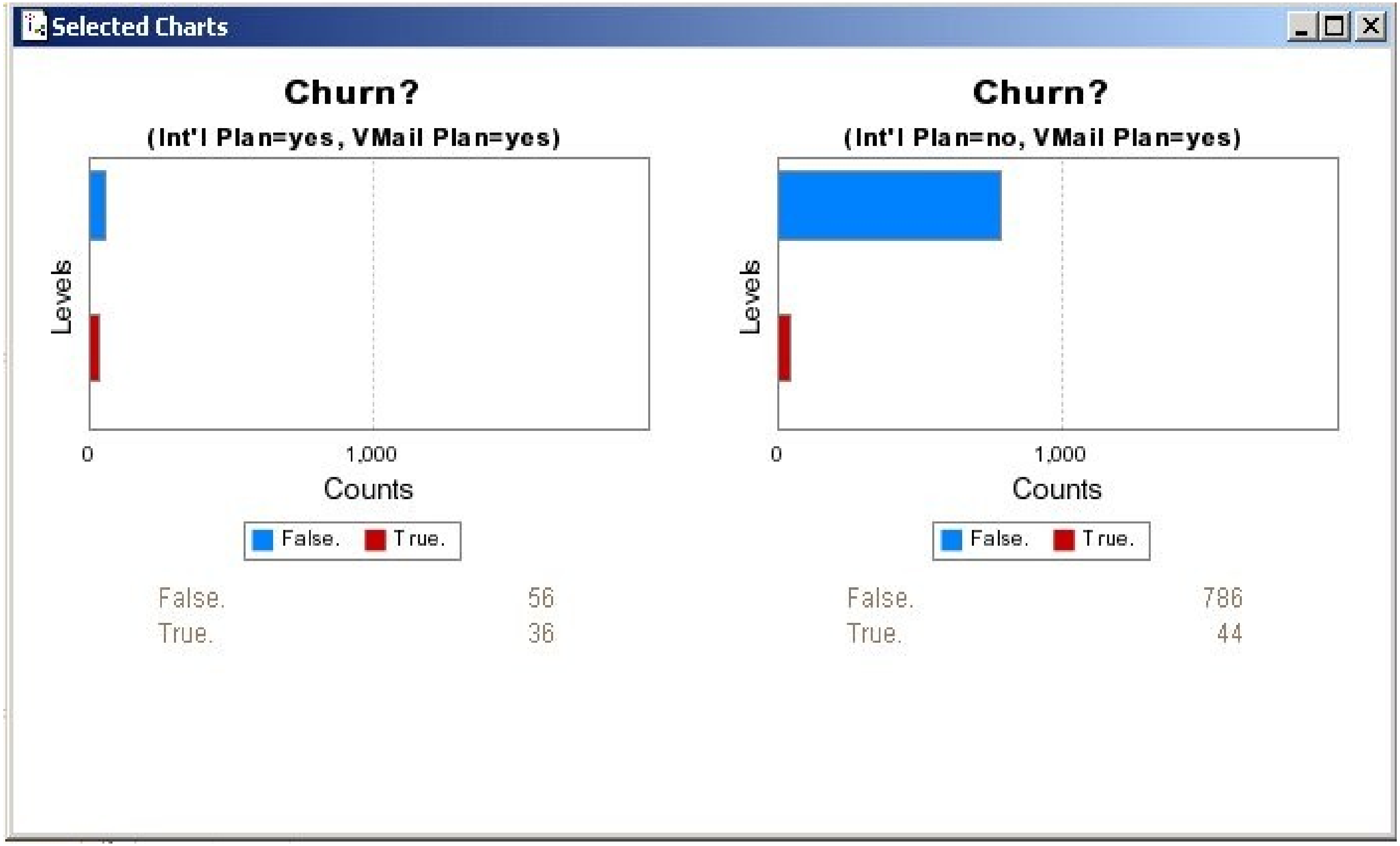




Comparing Multiple Variables



Comparing Multiple Variables



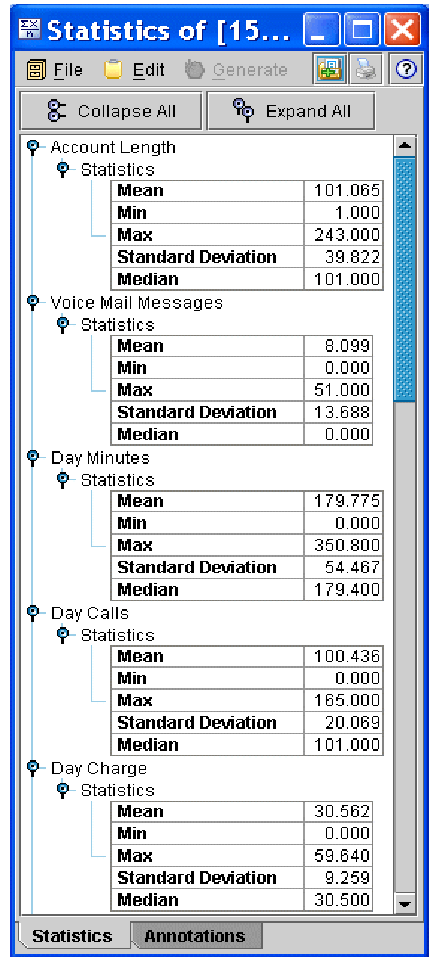
Summary of EDA for International Plan

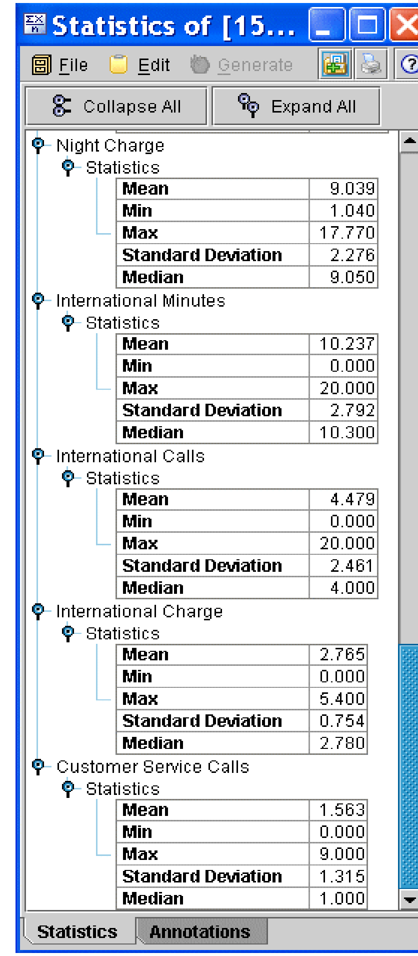
- Perhaps we should investigate what it is about our international plan that is inducing our customers to leave

- We should expect that, whatever data mining algorithms we use to predict churn, the model will probably include whether or not the customer selected the International Plan

Exploring Numeric Variables

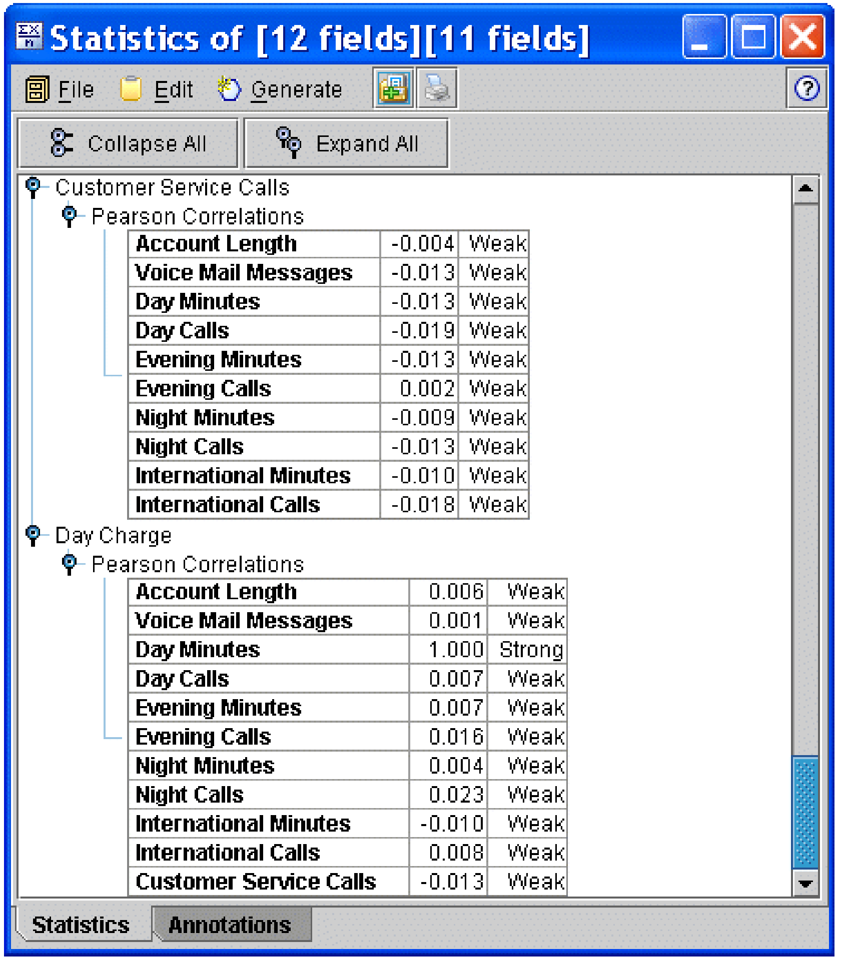
- Numeric summary measures for several variables shown
- Includes min and max, mean, median, and standard deviation
- For example, *Account Length* has min = 1 and max = 243
- Mean and median both ~101, which indicates symmetry
- *Voice Mail Messages* not symmetric; mean = 8.1 and median = 0





Exploring Numeric Variables *(cont'd)*

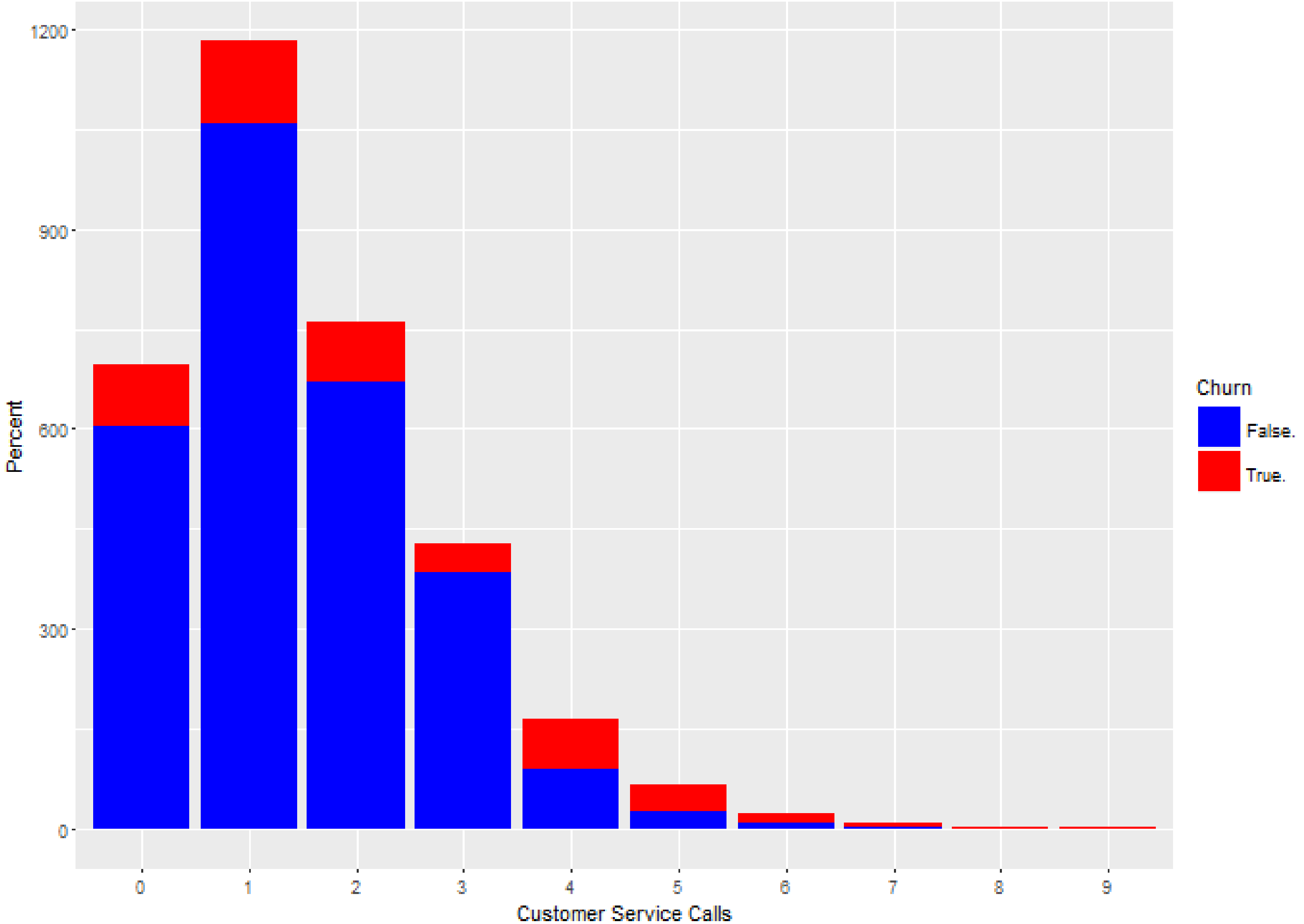
- Median = 0 indicates half of customers had no voice mail messages
- Recall use of correlated variables should be avoided
- Correlations of *Customer Service Calls* and *Day Charge* with other numeric variables shown
- All correlations are "Weak" except for *Day Charge* and *Day Minutes*, where *r* = 1.0
- Indicates perfect linear relationship



Account Length	-0.004	Weak
Voice Mail Messages	-0.013	Weak
Day Minutes	-0.013	Weak
Day Calls	-0.019	Weak
Evening Minutes	-0.013	Weak
Evening Calls	0.002	Weak
Night Minutes	-0.009	Weak
Night Calls	-0.013	Weak
International Minutes	-0.010	Weak
International Calls	-0.018	Weak

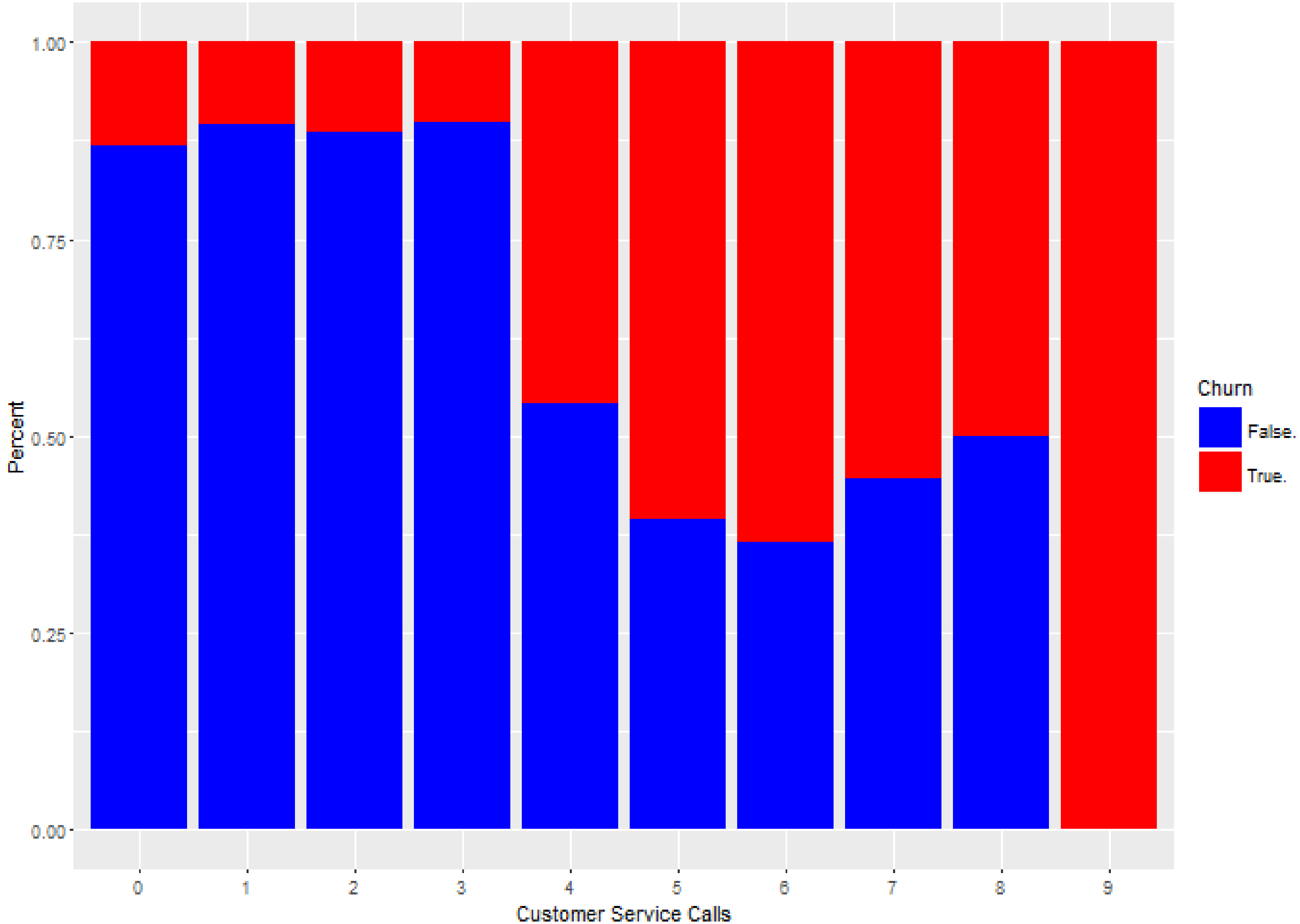
Account Length	0.006	Weak
Voice Mail Messages	0.001	Weak
Day Minutes	1.000	Strong
Day Calls	0.007	Weak
Evening Minutes	0.007	Weak
Evening Calls	0.016	Weak
Night Minutes	0.004	Weak
Night Calls	0.023	Weak
International Minutes	-0.010	Weak
International Calls	0.008	Weak
Customer Service Calls	-0.013	Weak

Histograms

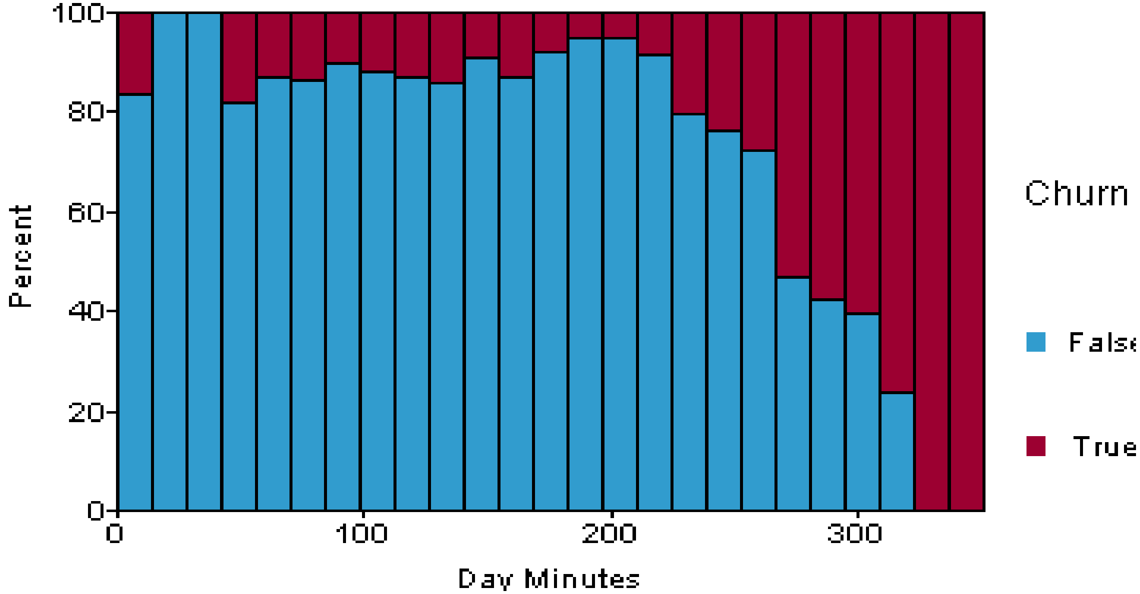


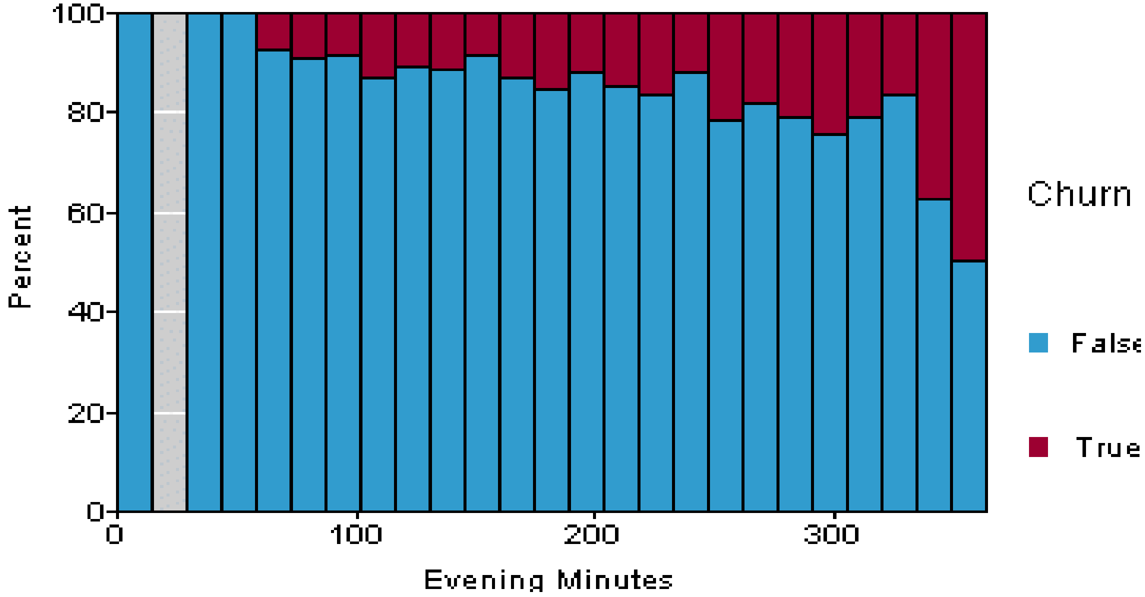
1200
900
600
300
0
Percent
0
1
2
3
4
5
6
7
8
9
Customer Service Calls
Churn
False.
True.

Normalized Histograms



Histograms *(cont'd)*



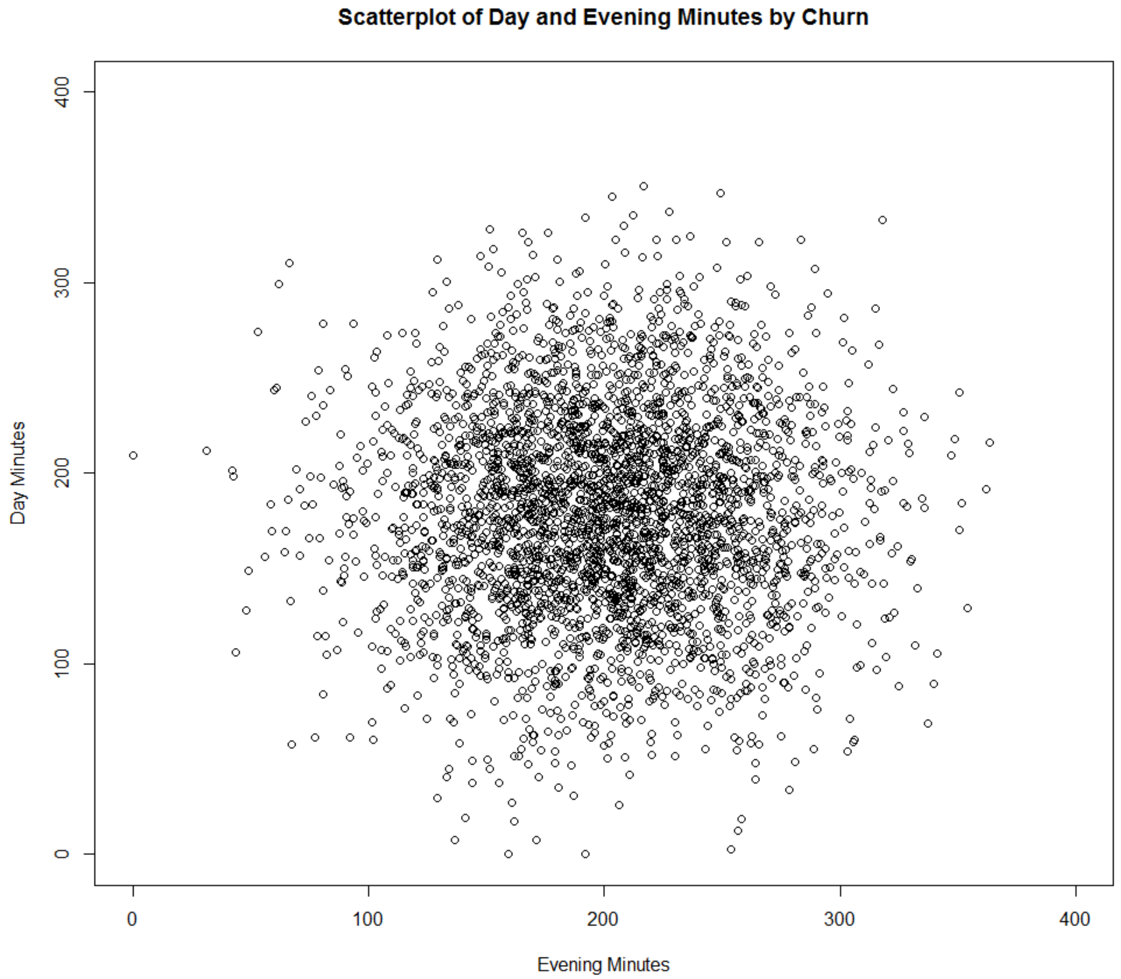


Summary of Additional Variables

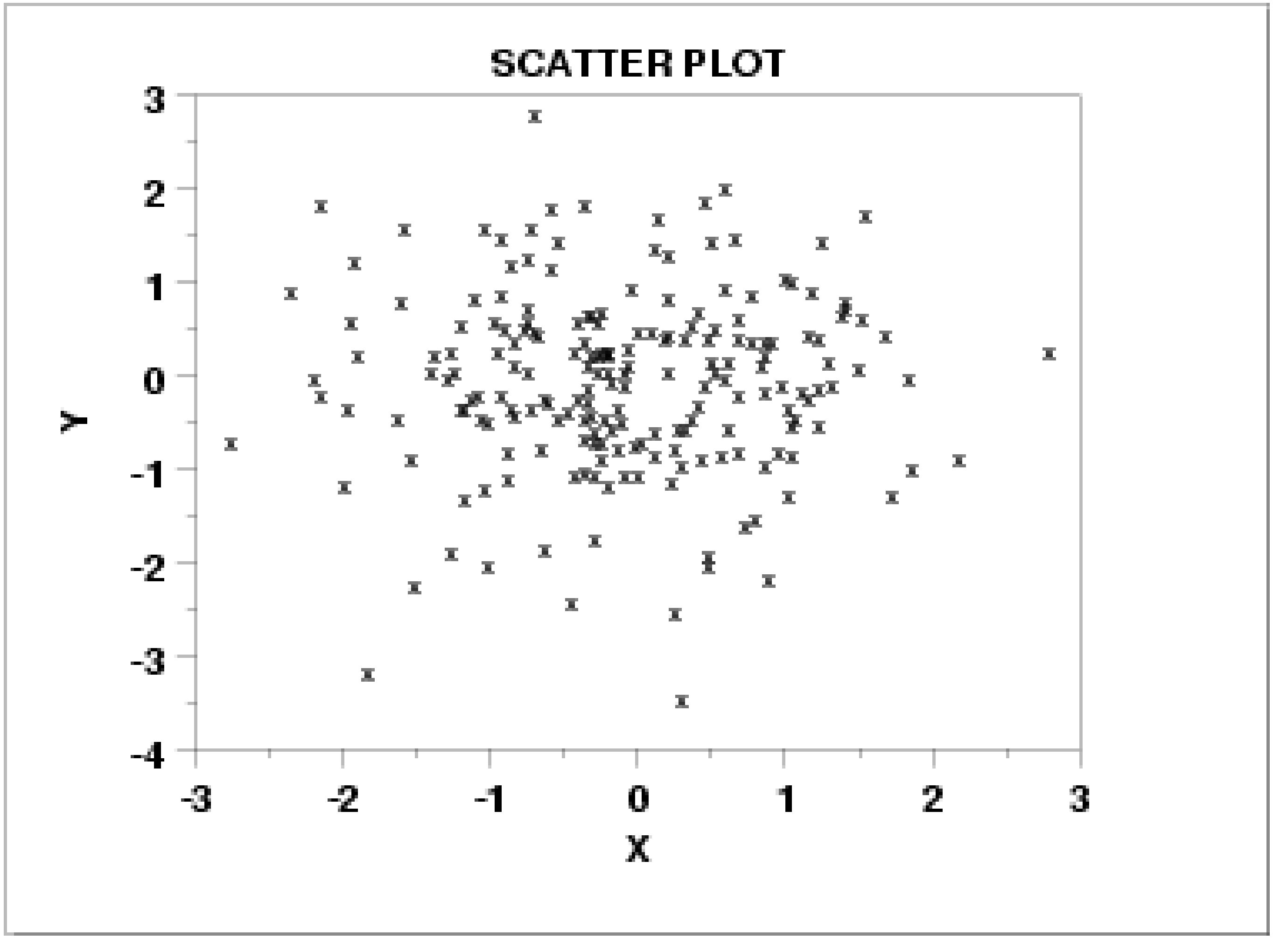
- Additional EDA concludes **no obvious association** between *Churn* and remaining numeric attributes
- These numeric attributes probably not strong predictors in data model; however, they should be retained as input to model
- Important higher-level associations/interactions may exist
- Let model identify which inputs are important
- Different EDA task may encounter huge number of inputs
- **Data mining performance adversely affected by many inputs?**
- Possibly exclude inputs not associated with target variable
- Or, use dimension-reduction technique such as principal components analysis.

Exploring Multiple Numeric Variables (Multivariate Relations)

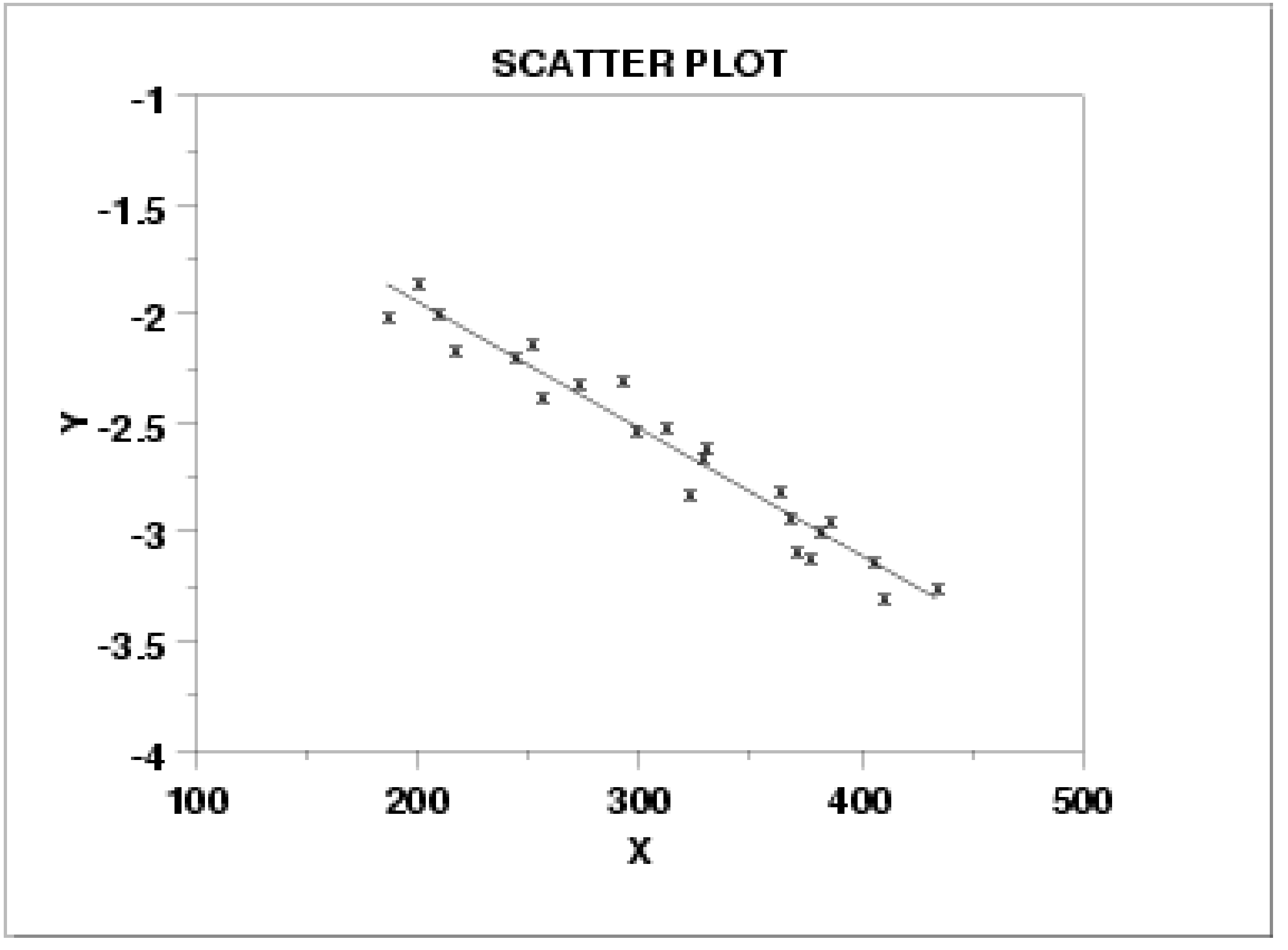
- Scatter Plots



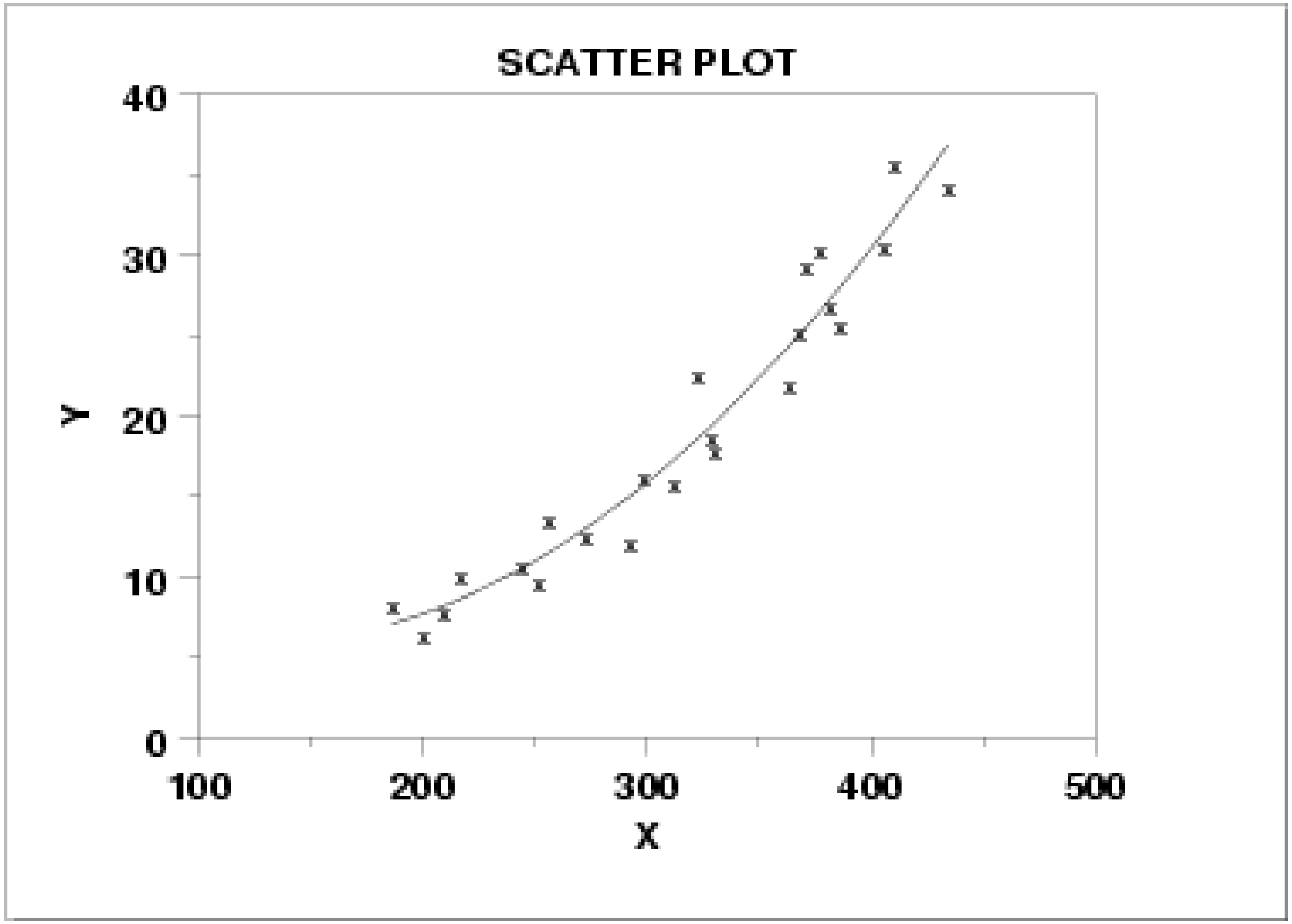
Scatter Plots: No apparent relationship



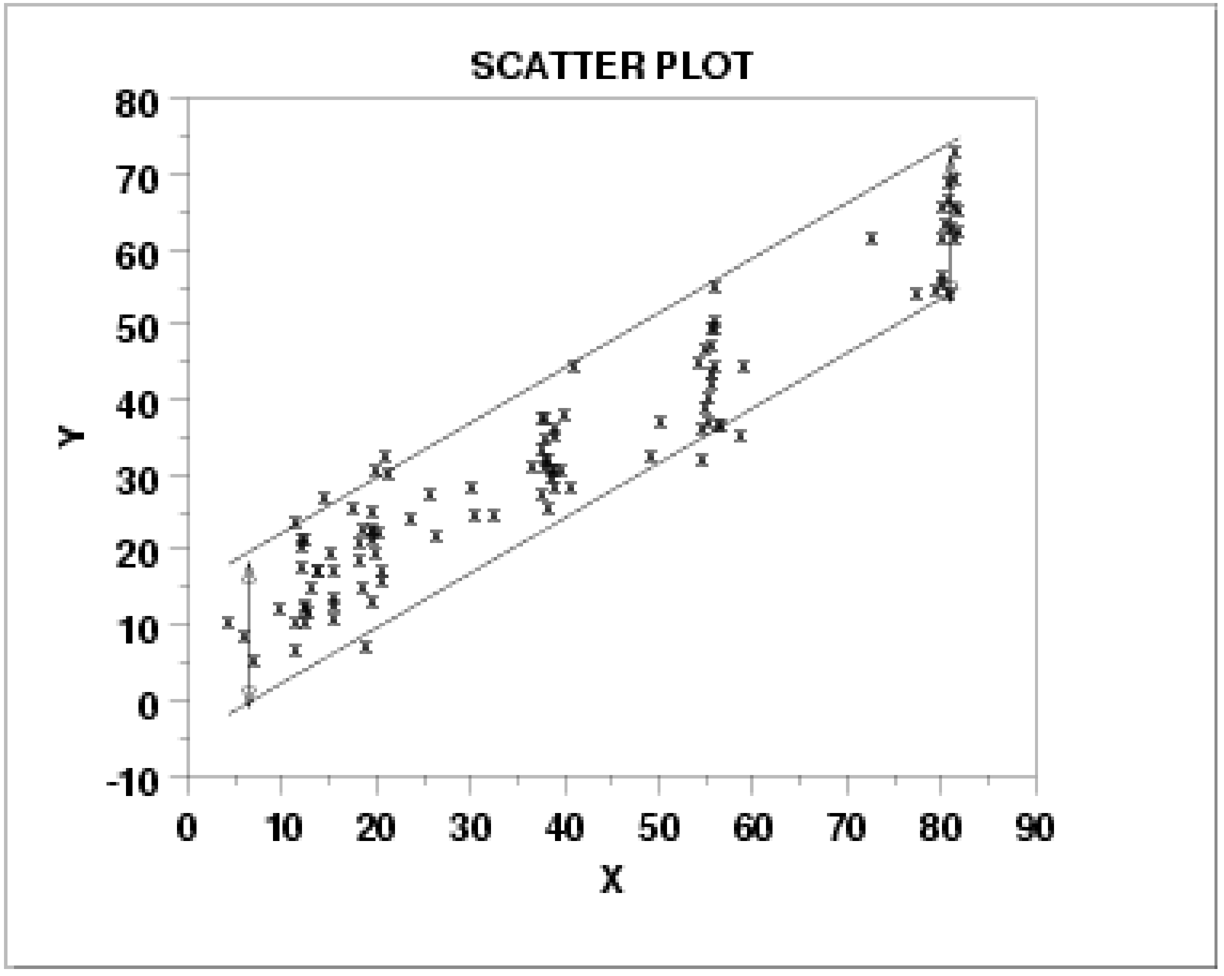
Scatter Plot: Linear Relationship



Scatter Plot: Quadratic Relationship

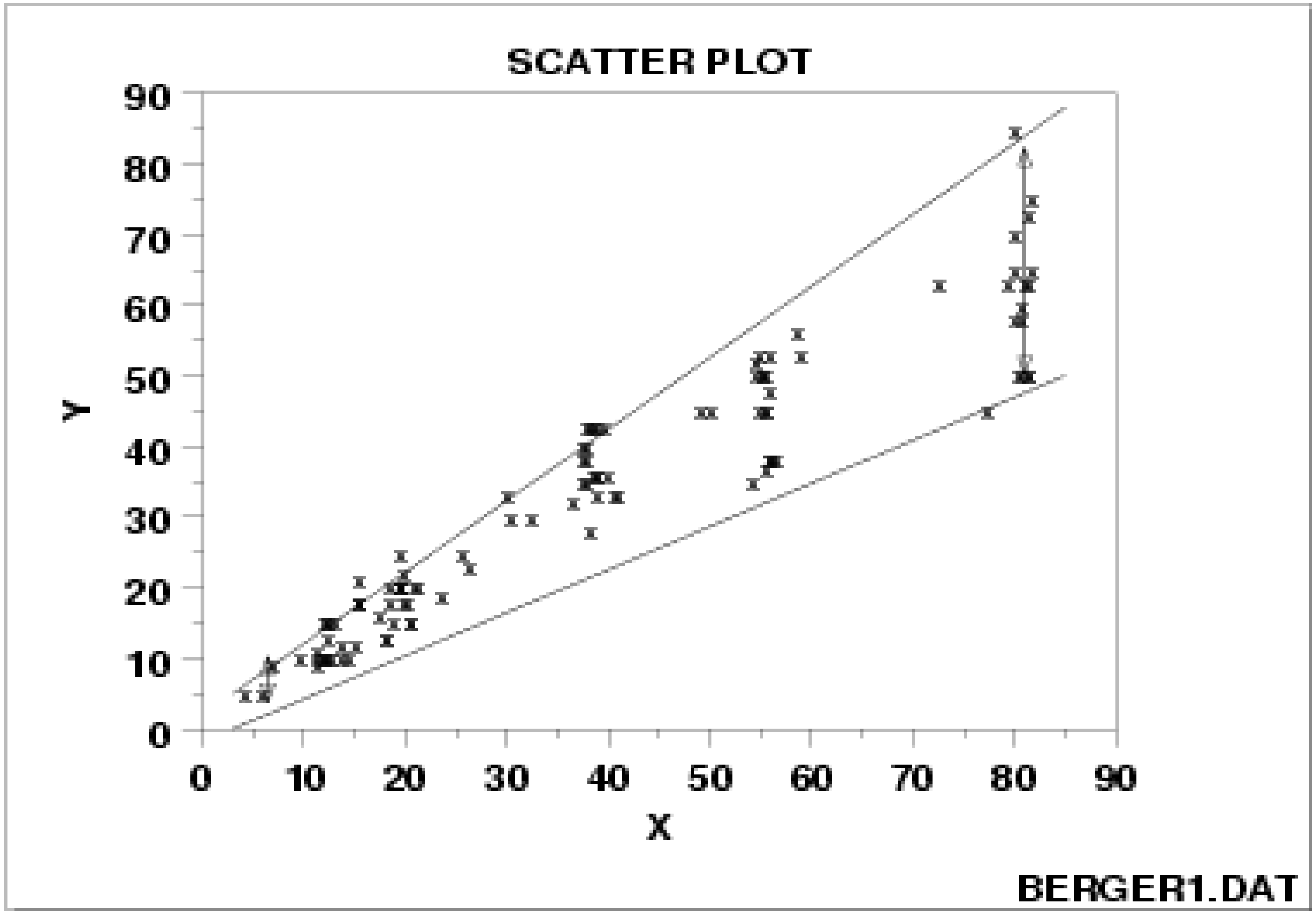


Scatter Plot: Homoscedastic



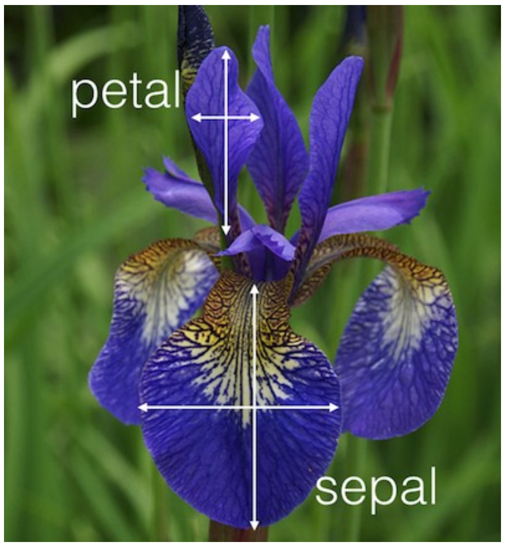
As x increases the variance of y does not change

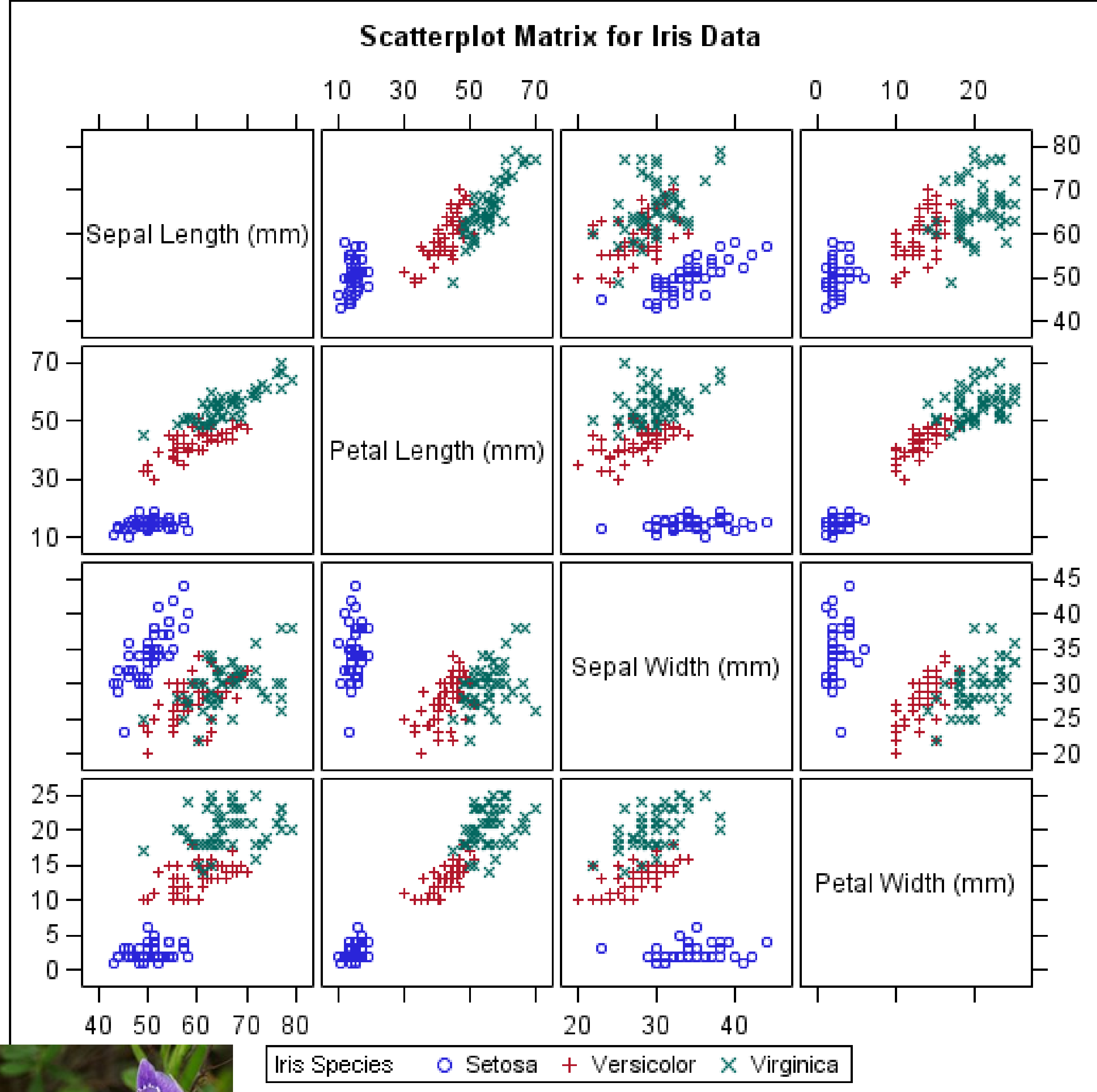
Scatter Plot: Heteroscedastic



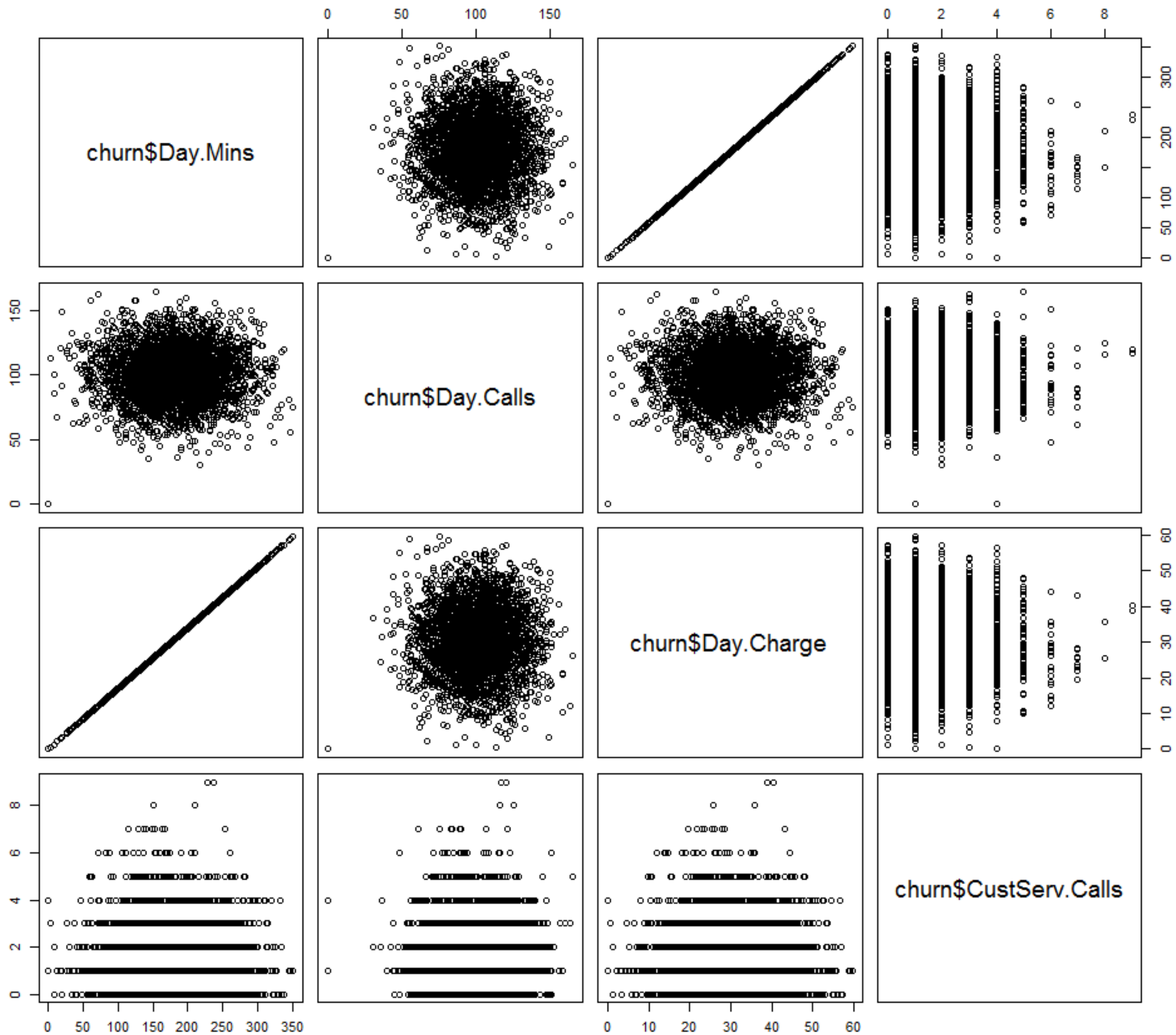
As x increases, the variance of y changes - in this case increases

More than two variables

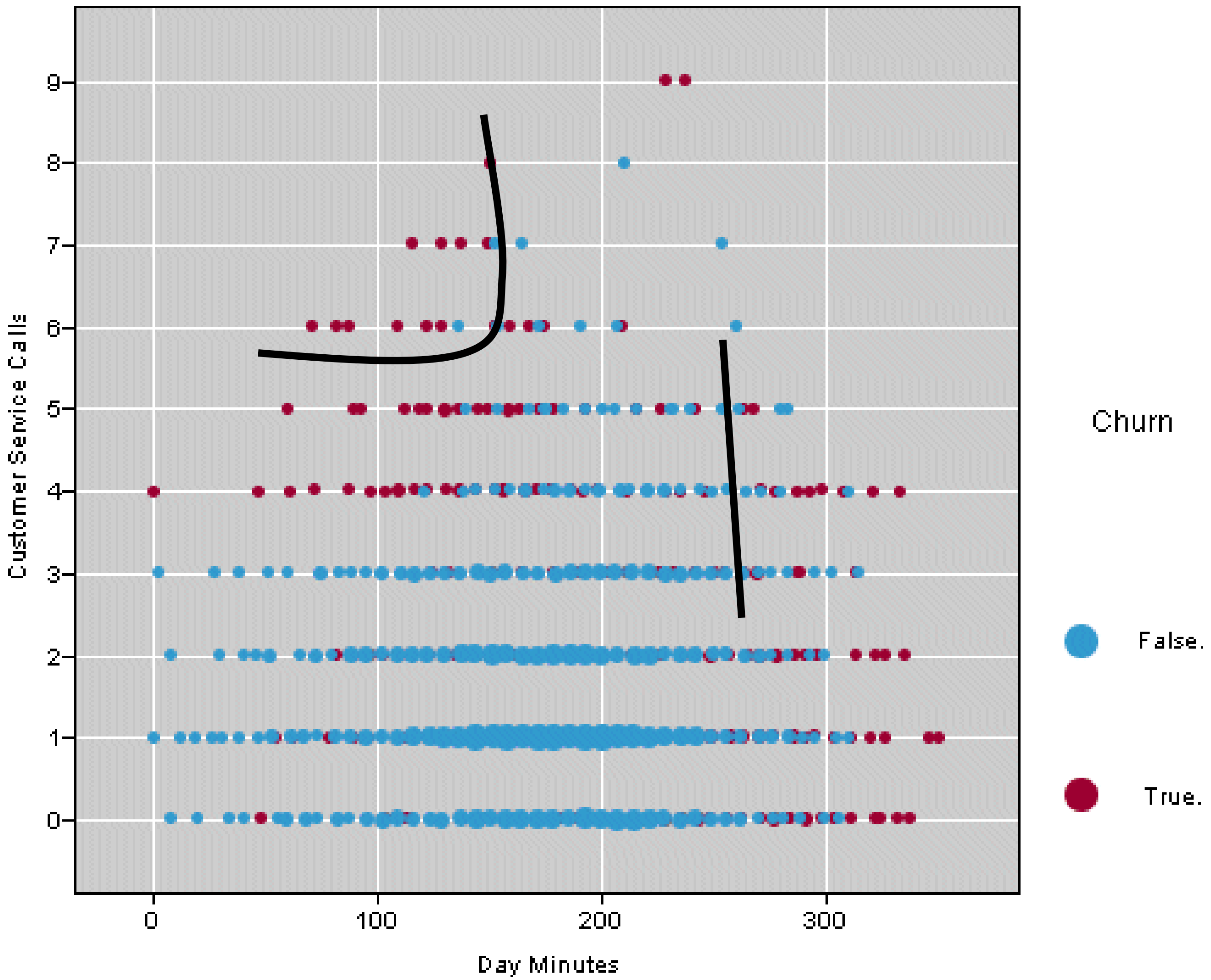




Scatter Plot Matrix of Day Minutes, Day Calls, Day Charge, and Customer Service Calls

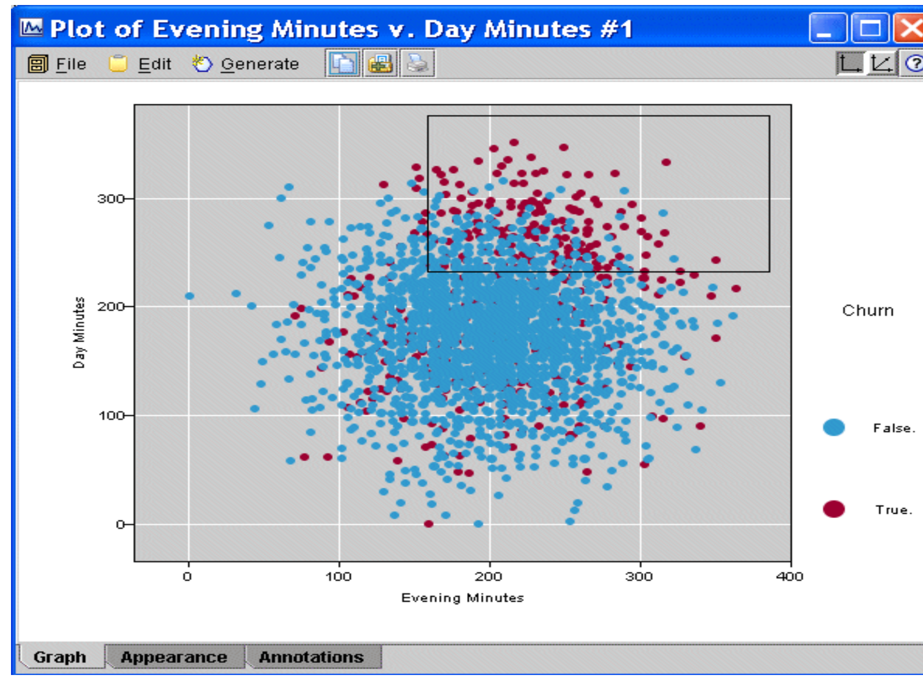


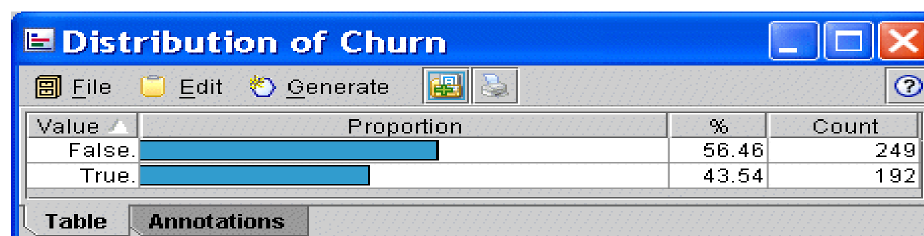
Scatter Plot of Day Minutes and Customer Service Calls Colored by Churn



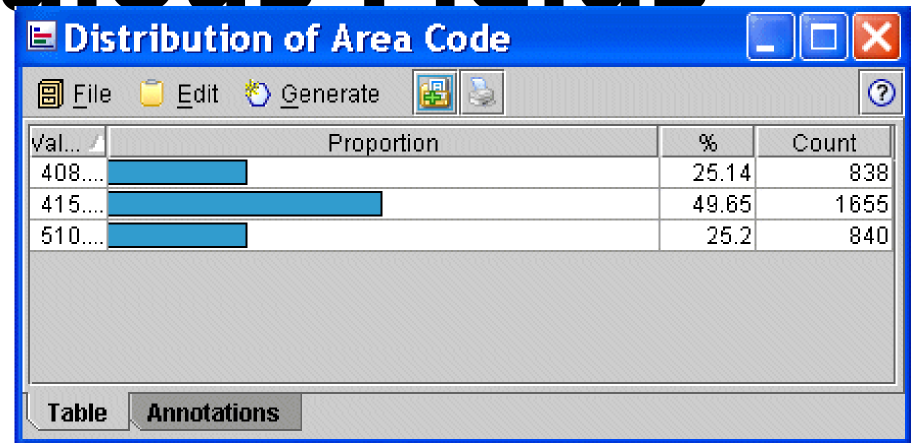
Selecting Interesting Subsets of the Data for Further Investigation

- Scatter plots or histograms identify interesting subsets of data
- Top figure shows selection of churners with high day and evening minutes
- Distribution of churn for this subset shown (bottom)
- 43.5% (192/441) of customers having both high day and evening minutes are churners
- This is ~3X churn rate of entire data set





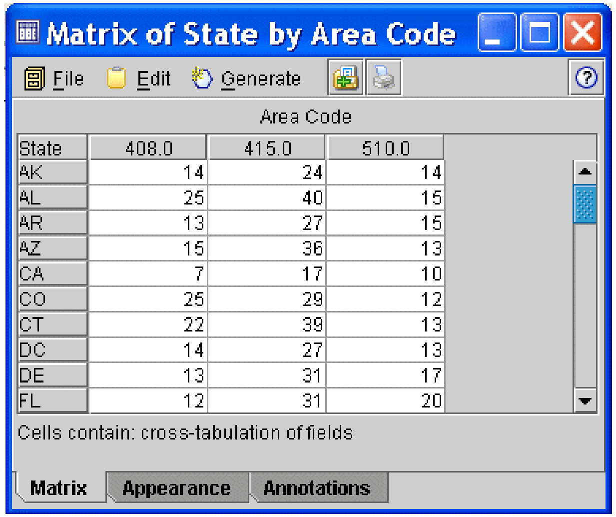
Using EDA to Uncover Anomalous Fields



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

- EDA sometimes uncovers anomalous records
- For example, examine distribution of *Area Code* variable
- *Area Code* used as categorical variable, grouping records geographically
- Attribute contains only three values: 408, 415, and 510
- All area codes located in California
- Is this strange?
- Perhaps not, if all records from California

Using EDA to Uncover Anomalous Fields *(cont'd)*



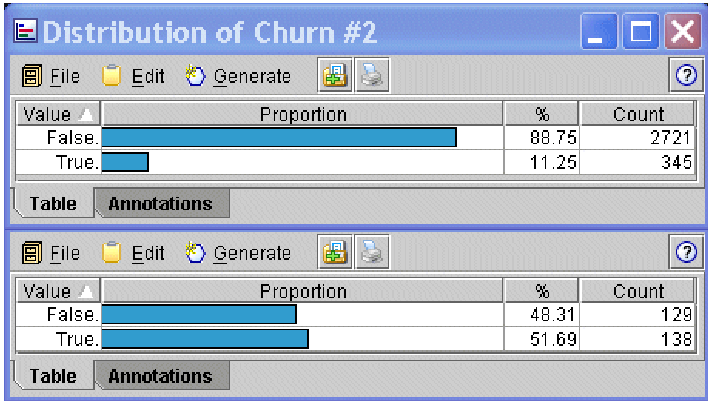
State	408.0	415.0	510.0
AK	14	24	14
AL	25	40	15
AR	13	27	15
AZ	15	36	13
CA	7	17	10
CO	25	29	12
CT	22	39	13
DC	14	27	13
DE	13	31	17
FL	12	31	20

- However, cross-tabulation of *Area Code* and *State* shows an anomaly
- Area codes distributed evenly across <u>all states</u>
- Data for attribute likely in error; or *State* attribute may have incorrect values?
- Domain expert should be consulted before including these variables in data mining models

Binning

- Binning categorizes an attribute's numeric (or categorical) values into reduced set of classes
- Makes analysis more convenient
- For example, number of *Day Minutes* could be binned into "Low", "Medium", and "High" categories
- For example, *State* values may be binned into regions
- California, Oregon, Washington, Alaska, and Hawaii are categorized as "Pacific"
- Binning defined as both data preparation and data exploration activity
- Various strategies exist for binning numeric variables
- One approach equalizes number of records in each class
- Another partitions values into groups, with respect to target

Binning *(cont'd)*



Value	Proportion	%	Count
False.		88.75	2721
True.		11.25	345

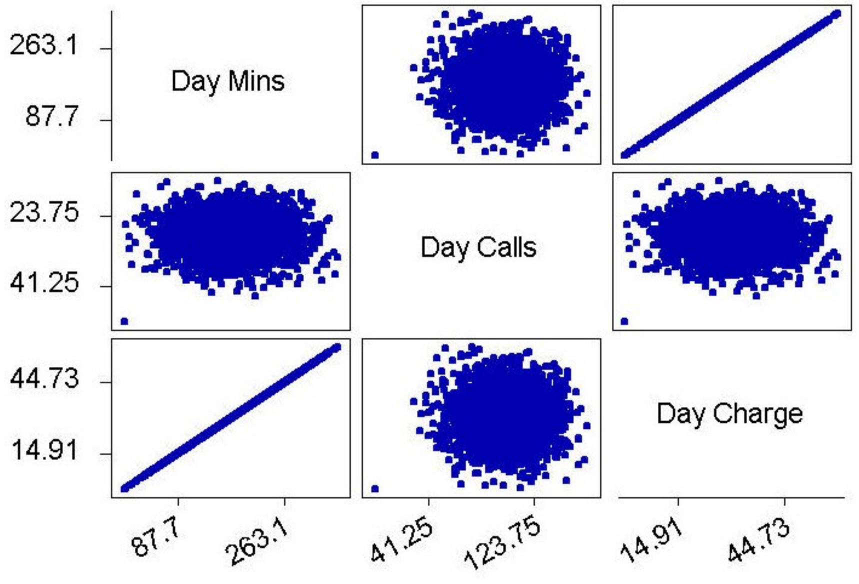
Value	Proportion	%	Count
False.		48.31	129
True.		51.69	138

- Recall those with fewer *Customer Service Calls* have lower churn rate
- For example, bin number of *Customer Service Calls* into "low" and "high" categories
- Figure shows churn rate for "low" class is 11.25% (Top)
- However, those within "high" group have 51.69% churn rate (Bottom)
- Churn rate more than 4X higher

Dealing with Correlated Variables

- Using highly correlated variables in data model:
  - Should be avoided!
  - Incorrectly emphasizes one or more data inputs
  - Creates model instability and produces unreliable results

  - Matrix plot of *Day Minutes*, *Day Calls*, and *Day Charge* shown in



Dealing with Correlated Variables *(cont'd)*

- As number of *Day Minutes* increase we expect *Day Charge* to increase
- Example of <u>positive correlation</u>
- Oddly, lack of graphical evidence supports correlation between *Day Minutes* and *Day Calls*, or *Day Calls* and *Day Charge*
- Additionally, $r = 0.07$ indicating variables uncorrelated

- However, <u>linear relationship</u> exists between *Day Charge* and *Day Minutes*
- *Day Charge* is <u>linear function</u> of *Day Minutes*

Strategy for Handling Correlated Variables

- Identify any variables that are perfectly corrected
  - Omit one.

- Identify groups of variables that are correlated with each other
  - Apply dimension reduction methods during the modeling phase

Dealing with Correlated Variables *(cont'd)*

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%
```

- Estimated regression equation shown in Figure 3.3 (Minitab) expresses relationship

  "Day Charge equals 0.000613 plus 0.17 times Day Minutes"
- Company uses flat-rate billing model of 17 cents/minute
- *R*-squared statistic = 1.0 ☾ indicates perfect linear relationship
- Therefore, *Day Charge* and *Day Minutes* are correlated

Dealing with Correlated Variables *(cont'd)*

- One of two variables should be eliminated from model
- *Day Charge* arbitrarily chosen for removal
- *Evening*, *Night*, and *International* variable pairs reflect similar results
- Therefore, *Evening Charge*, *Night Charge*, and *International Charge* also removed
- Proceeding to data mining without first eliminating correlated variables may have produced compromised results
- Number of attributes reduced from 20 to 16
- Reduction in dimensionality of solution space beneficial to some data mining algorithms

Summary

- EDA uncovered some insights into *churn* data set:
  - Four “Charge” fields are linear functions of “Minutes” fields
  - Correlation among remaining numeric attributes “Weak”
  - *Area Code* and/or *State* fields anomalous
  - Customers with *International Plan* churn at higher rate
  - Those in *Voice Mail Plan* churn less frequently
  - Customers calling customer service 4 or more churn 4X higher than others
  - Customer with high day and evening minutes churn 4X higher rate than others
- These observations performed using EDA only; no data mining applied
- Results can be easily formulated into actionable plan designed to reduce churn rate
- *Useful links:*
  - *https://r4ds.had.co.nz/exploratory-data-analysis.html*
  - *https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce*
  - *https://www.kaggle.com/ekami66/detailed-exploratory-data-analysis-with*