Chapter 3:
Exploratory Data Analysis

Introduction
Exploring Categorical Variables
Exploring Numerical Variables
Exploring Categorical/Numerical Variables
Selecting Interesting Subsets of Data
Binning

Review

- Data mining tasks
  - Description (Ch. 3: Exploratory Data Analysis)
  - Estimation (Ch. 4: Statistical Approaches)
  - Classification (Ch. 5, 6 & 7: k-NN, DT & NN)
  - Prediction (Chapters 4 to 7)
  - Clustering (Ch. 8 & 9: k-means & Kohonen Networks)
  - Association (Ch. 10)

Hypothesis Testing vs. EDA

- hypothesis-testing is used when there is a priori hypotheses
  - Example: whether a recent increase in the cell phone fee has led to a decrease in market share
- Otherwise, EDA or *graphical data analysis is used*
  - Especially, when there is large unknown databases.

3

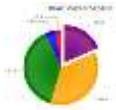
What EDA can do

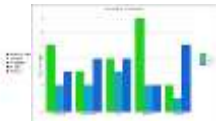
- Delve into the data set
- Examine the interrelationships among the attributes
- Identify interesting subsets of the observations
- Develop an initial idea of possible associations between the attributes and the target variable, if any

4

How?

- Three combinations:
  - 2 Categorical variables (gender, race, state,…):
    - Pie chart, bar chart, …
    - Cross Tabulation
    - …





  - 2 Numerical variables (height, age, salary, price,…):
    - Regression or Correlation
    - Histogram
    - Analyzing mean, median, standard dev., …
    - Scatter Plot
    - …
  - 1 Categorical, 1 Numerical variables
    - Side-by-side box plots

5

Churn dataset

Table (21 fields, 3,333 records) #1

File Edit Generate

	State	Account Length	Area Code	Phone	Intl Plan	VMail Plan	VMail Messages	Day Mins
1	KS	128	415	382-4657	no	yes	25	265.100
2	OH	107	415	371-7191	no	yes	26	161.600
3	NJ	137	415	358-1921	no	no	0	243.400
4	OH	84	408	375-9999	yes	no	0	299.400
5	OK	75	415	330-6626	yes	no	0	166.700
6	AL	118	510	391-8027	yes	no	0	223.400
7	MA	121	510	355-9993	no	yes	24	218.200
8	MO	147	415	329-9001	yes	no	0	157.000
9	LA	117	408	335-4719	no	no	0	184.500
10	WV	141	415	330-8173	yes	yes	37	258.600

6

Churn dataset

Table (21 fields, 3,333 records) #1

File Edit Generate

	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl
1	110	45.070	197.400	99	16.780	244.700	91	11.010	1
2	123	27.470	195.500	103	16.620	254.400	103	11.450	1
3	114	41.380	121.200	110	10.300	162.600	104	7.320	1
4	71	50.900	61.900	88	5.260	196.900	89	8.860	
5	113	28.340	148.300	122	12.610	186.900	121	8.410	1
6	90	37.900	220.600	101	10.750	203.900	110	9.100	
7	88	37.090	348.500	108	29.620	212.600	118	9.570	
8	79	26.690	103.100	94	8.760	211.800	96	9.530	
9	97	31.370	351.600	80	29.890	215.800	90	9.710	
10	84	43.960	222.000	111	18.870	326.400	97	14.690	1

7

Churn dataset

Table (21 fields, 3,333 records) #1

File Edit Generate

	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls	Churn?
1	244.700	91	11.010	10.000	3	2.700	1	False.
2	254.400	103	11.450	13.700	3	3.700	1	False.
3	162.600	104	7.320	12.200	5	3.290	0	False.
4	196.900	89	8.860	6.600	7	1.780	2	False.
5	186.900	121	8.410	10.100	3	2.730	3	False.
6	203.900	118	9.180	6.300	6	1.700	0	False.
7	212.600	118	9.570	7.500	7	2.030	3	False.
8	211.800	96	9.530	7.100	6	1.920	0	False.
9	215.900	90	9.710	9.700	4	2.350	1	False.
10	326.400	97	14.690	11.200	5	3.020	0	False.

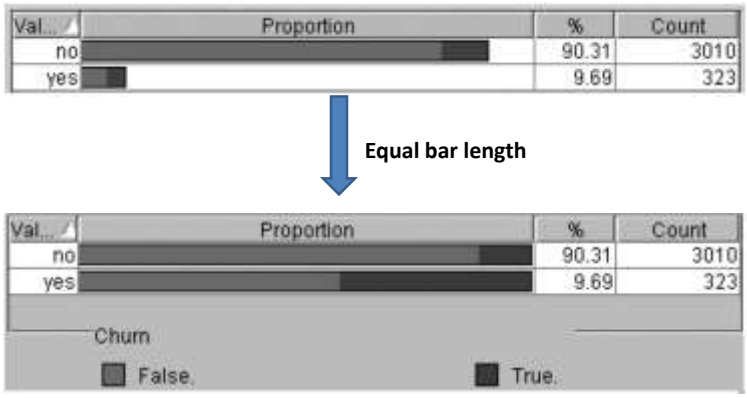
8

Review: two types of data

- Categorical variables (Qualitative)
- Numerical variables (Quantitative)

9

Exploring Categorical Variables:
Bar chart: international plan



Those who have selected the International Plan have a greater chance of leaving the company's service

10

Exploring Categorical Variables:
Cross Tabulation: international plan

To quantify the relationship between International Plan and churning:

	International Plan	
Churn	no	yes
False.	2664	186
True.	346	137

Cells contain: cross-tabulation of fields

Total churned: $\frac{346 + 137}{346 + 137 + 2664 + 186} = 14.5\%$

$346/(346 + 2664) = 11.5\%$ Of total churned, did not use Int. Plan

$137/(137 + 186) = 42.4\%$ Of total churned, used Int. Plan

11

Exploring Categorical Variables:
Cross Tabulation: international plan

1. What it is about the International Plan that is inducing customers to leave!
2. The data mining model we use to predict churn will probably include whether or not the customer selected the International Plan.

12

Uncover Anomalous Fields

	Area Code		
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

Cross Tabulation of Area Code and State

Anomaly: three area codes distributed across all 50 states.

=> Eliminating the redundant field

13

Exploring Numerical Variable: statistical parameters

Account Length	
Statistics	
Mean	101.065
Min	1.000
Max	243.000
Standard Deviation	39.822
Median	101.000

Voice Mail Messages	
Statistics	
Mean	8.099
Min	0.000
Max	51.000
Standard Deviation	13.688
Median	0.000

Day Minutes	
Statistics	
Mean	179.775
Min	0.000
Max	350.800
Standard Deviation	54.467
Median	179.400

Night Charge	
Statistics	
Mean	9.039
Min	1.040
Max	17.770
Standard Deviation	2.276
Median	9.050

International Minutes	
Statistics	
Mean	10.237
Min	0.000
Max	20.000
Standard Deviation	2.792
Median	10.300

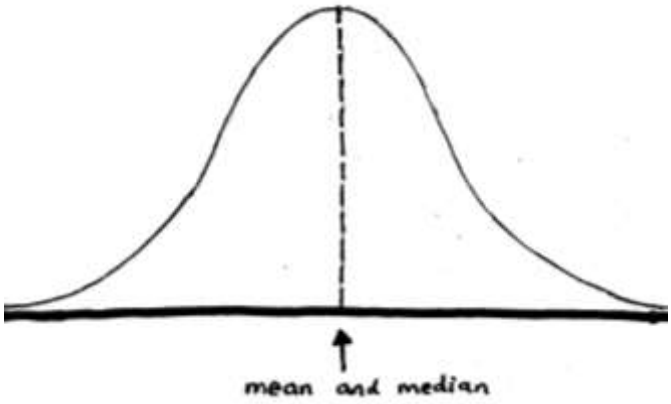
International Calls	
Statistics	
Mean	4.479
Min	0.000
Max	20.000
Standard Deviation	2.461
Median	4.000

Compare mean and median in each field

14

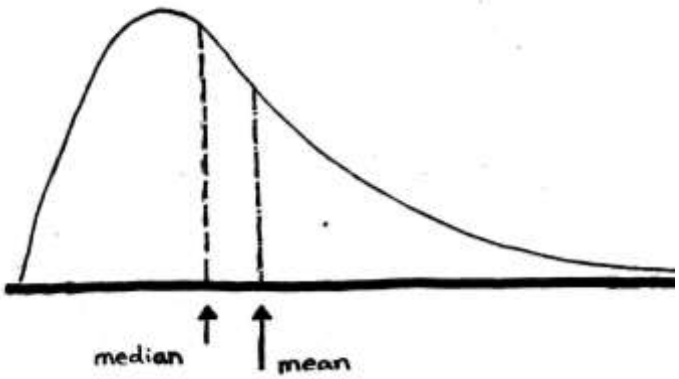
Mean vs Median

- Symmetric distribution



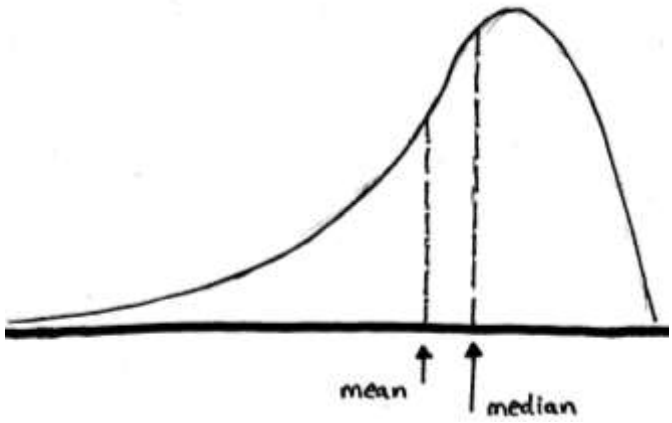
Mean vs Median

- Right skewed distribution



Mean vs Median

- Left skewed distribution



Extreme example

- Income in small town of 6 people

  $25,000 $27,000 $29,000

  $35,000 $37,000 $38,000

- Mean is $31,830 and median is $32,000
- Bill Gates moves to town

  $25,000 $27,000 $29,000

  $35,000 $37,000 $38,000 $40,000,000

- Mean is $5,741,571 median is $35,000
- Mean is pulled by the outlier while the median is not. The median is a better of measure of center for these data

Exploring Numerical Variable: Correlation

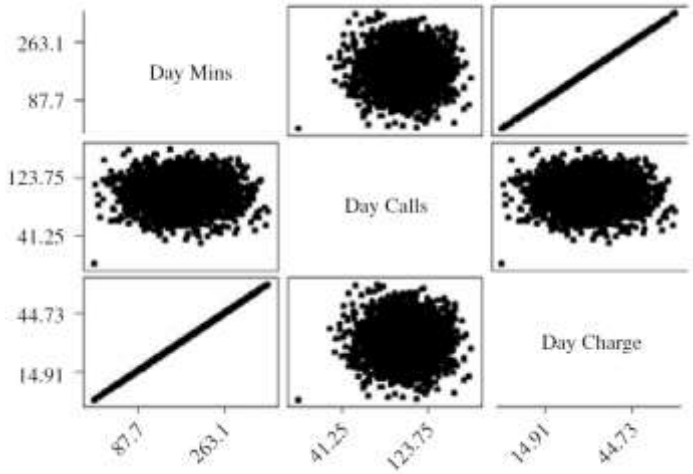
Day Charge

Pearson Correlations

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

19

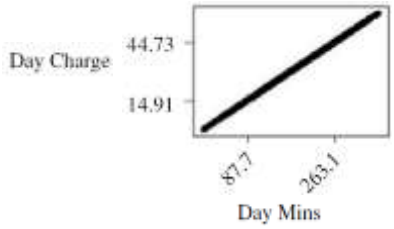
Dealing with Correlated variables



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

20

Dealing with Correlated variables

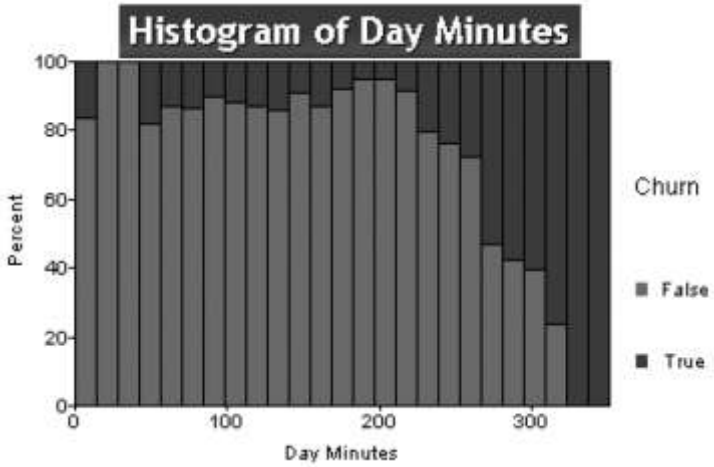



```
The regression equation is
Day Charge = 0.000613 + 0.170 Day Mins
```

=> Eliminating redundant variable

21

Exploring Numerical Variable: Histogram



22

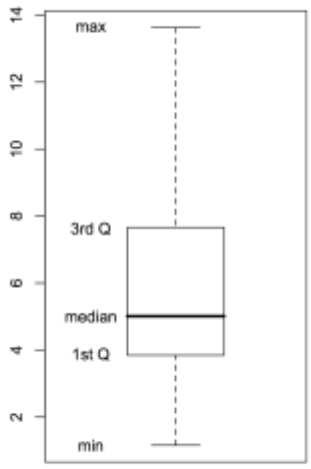
Summary of one variable EDA

Variable	Disposition
State	Anomalous. Omitted from model.
Account length	No obvious relation with churn, but retained.
Area code	Anomalous. Omitted from model.
Phone number	Surrogate for ID. Omitted from model.
International Plan	Predictive of churn. Retained.
VoiceMail Plan	Predictive of churn. Retained.
Number of voice mail messages	No obvious relation with churn, but retained.
Total day minutes	Predictive of churn. Retained.
Total day calls	No obvious relation with churn, but retained.
Total day charge	Function of *minutes*. Omitted from model.
Total evening minutes	May be predictive of churn. Retained.
Total evening calls	No obvious relation with churn, but retained.
Total evening charge	Function of *minutes*. Omitted from model.
Total night minutes	No obvious relation with churn, but retained.
Total night calls	No obvious relation with churn, but retained.
Total night charge	Function of *minutes*. Omitted from model.
Total international minutes	No obvious relation with churn, but retained.
Total international calls	No obvious relation with churn, but retained.
Total international charge	Function of *minutes*. Omitted from model.
Customer service calls	Predictive of churn. Retained.

23

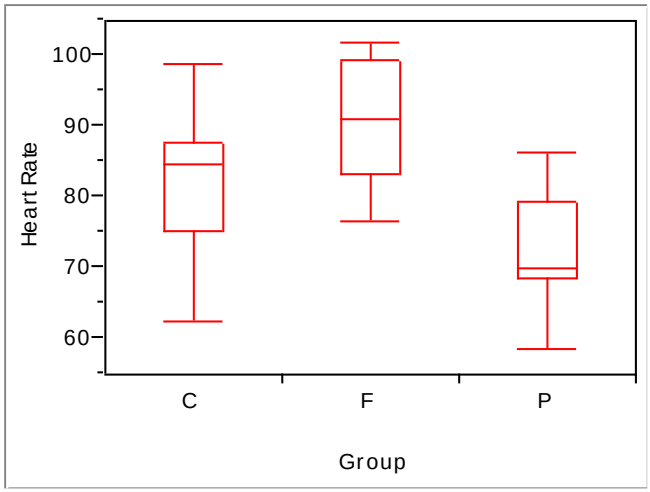
EDA: 1 Categorical, 1 Numerical variables: Box plots

- A box plot is a graph of five numbers (often called the five number summary)
  - minimum
  - Maximum
  - Median
  - 1st quartile
  - 3rd quartile

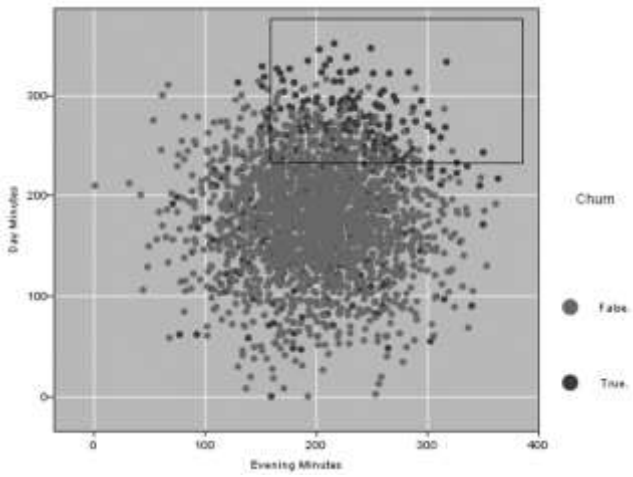


EDA: 1 Categorical, 1 Numerical variables: Box plots

Pets and Stress, Allen et al. (1988) **. Three groups: alone, Friend, Pet**



Selecting Interesting Subsets of the Data



43% of the subset are churners

Binning (or banding)

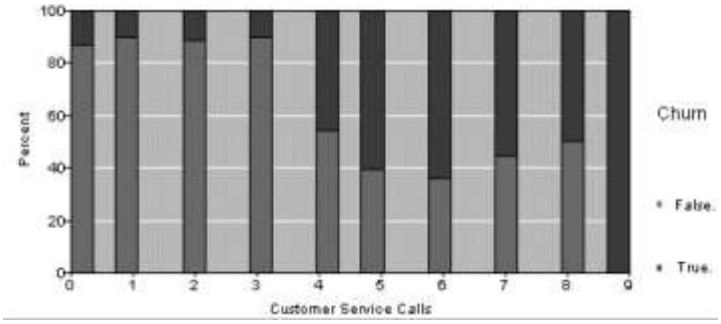
- Binning refers to the categorization of numerical or categorical variables into a manageable set of classes which are convenient for analysis.
- Example:
  - the number of day minutes could be categorized (binned) into three classes:
    - Low
    - Medium
    - high

27

Binning

1. the classes of equal width
2. equalizing the number of records in each class
3. identifiable groups of records, which, with respect to the target variable, have similar behavior
4. clustering

28



If Customer service call > 3,
class = High
Else
class = low

29