Data Mining

Lecture 2
Data Understanding

Lecturer: G. Gray

Data Understanding

Data Understanding activities

- Collecting the data
- Describing the data
- Doing an initial survey
- Verifying the quality of the data

(from CRISP-DM)

Data Understanding

- Different types of data found in data sets
- Need to understand the data to enable an assessment of the quality and information content of a data set

- The data understanding phase starts with an initial data collection and then proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover insights into the data, or to detect interesting subsets to form hypotheses for hidden information

Collecting the data

Collecting the data

- To perform any analysis on data it must be organised into a set of attributes, (as in a database table)
- Data mining algorithms work on a **single** data set which is a collection of records
  - For each mining objective, appropriate attributes need to be merged into a single table
- However accessing live databases directly is not practical as this data is continually being updated by transaction processing systems

- A **Data warehouse** is the most common source of data for a data mining project in business.

ethics

Access Issues

- **Legal issues**
  - E.G. Confidentiality rights for medical data
  - E.G. Can not hold credit information about someone to whom credit will not be offered
- **Departmental issues**
  - For ethical reasons one department may hide data from other departments (financial trading, salary details)
- **Political reasons**
  - Owner of the data many not be in favour of the data mining project
- **Data Format**
  - Media format (magnetic tape, diskettes etc.)
  - Format differences (ASCII, EBCDIC, binary packed decimal)
- **Connectivity**
  - Is the data available on-line
- **Architectural reasons**
  - Data from various types of databases – heterogeneous data sources
- **Timing**
  - Various data streams may not be equally current

technical

personalities

Describing the data

Properties of a data set

- Data types – is the data numerical or text/ alphanumeric?

- Dimensionality - number of attributes that are in database (no. of columns)
- Instances – number of rows in the dataset
  - Typically need at least **20s time** the numbers of rows as columns to represent all patterns.

- Resolution – any column can be calculated at different resolutions e.g. height could be recorded in centimeters or meters

Types of Attributes

t
e
x
t

- Nominal
  - Allocating names to things e.g. names, ID numbers
- Categorical
  - Names groups of things, arbitrary labels with no specific order e.g. eye colour, gender, zip code
- Ordinal
  - Labels have specific order (alpha or numeric) e.g. {tall medium, short}, grades, ratings {1,2,3}

---

n
u
m
e
r
i
c

- Interval
  - Has an order but also carries the information about the distance between values on the scale e.g. temperature (Celsius or Fahrenheit) calendar dates
- Ratio
  - As for interval but you can express meaningful ratios based on the numbers in the scale e.g. length, time counts,

Ratio V Interval

Height (inches)
60
50
40
30
20
10
Absolutely no height ← 0
This is the true zero point

Temperature °C
60
50
40
30
20
10
0 → *Some heat is present at 0*
- 10
- 20
- 30
.
.
.
- 273.15 → *Absolutely no heat*
This is the true zero point

Height (Inches)
60'': 30''
2: 1

Temperature °C
60 °C: 30 °C
60 + 273.15: 30 + 273.15
333.15: 303.15
1.1: 1

Types of Attributes

- Attribute can have a high or low information content

Lowest information content

- Names (nominal scale)
- Zip codes (categorical scale)
- Hot, warm, cool, cold (ordinal scale)
- -40°c, . ., 0°c, . . +40°c (interval scale)
- Bank balance (0,. . , 5,. . ,10, . ..) (ratio scale)

Highest information content

Exercise

For each of the following, state if it is nominal, categorical, ordinal, interval or ratio

- GPA
- US states
- Academic Grade
- Street name
- Area code (telephones)
- Account number
- Day charge (daily phone charge)
- Percentage grade

Numeric variables can be:

- Discrete attribute
  - Has only a finite set of values
  - E.g. Zip codes
  - Often represented as integer values

- Continuous
  - has real numbers as attributes values
  - E.g. temperature, height, weight
  - Practically, real values can only be measured and represented using a finite number of digits

- Questions: Which would be easier to predict?

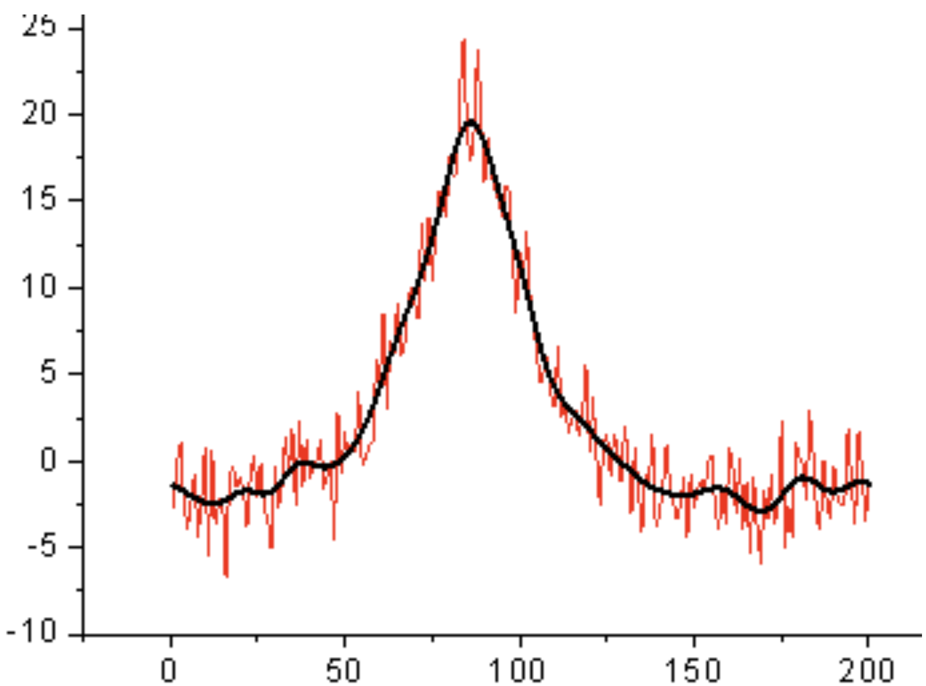
Verify data quality

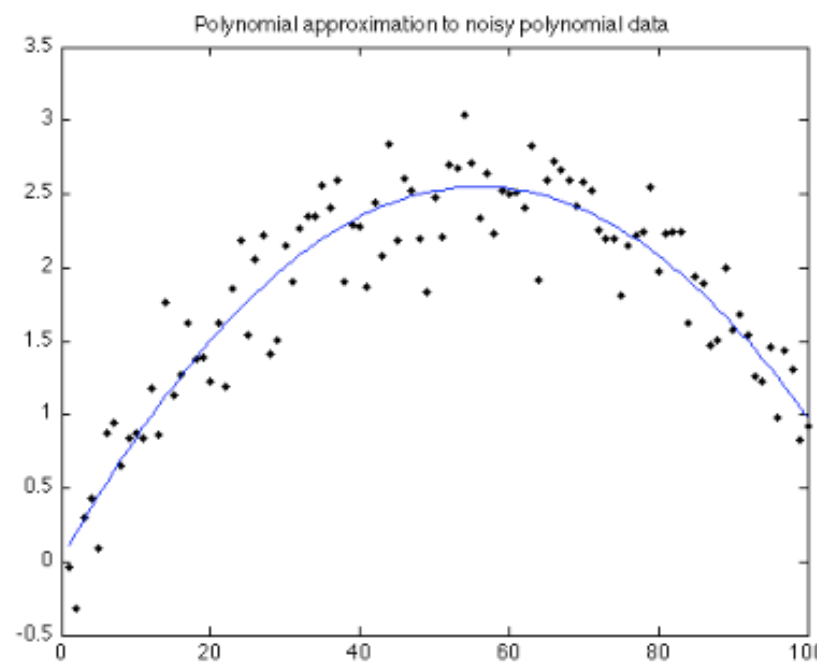
Data Quality

- Data quality issues
  - Noise
  - Outliers
  - Missing data
  - Inconsistent data
  - Duplicate data
  - Pollution in the data
  - Bias in the data

Noisy data

- Noise – any random error or variance in a variable
  - It is difficult to detect, but it may be visible as a greater variance in a variable's values or an outlier value





Outliers

- Outliers:
  - a value outside the expected range for that attribute
  - A combination of values that is unusual
- Examples:

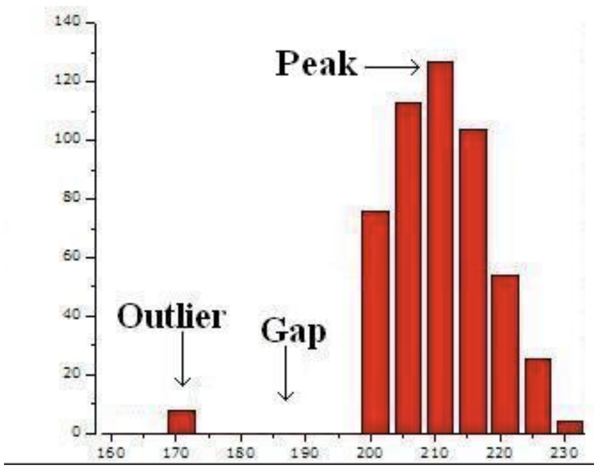
age = 106

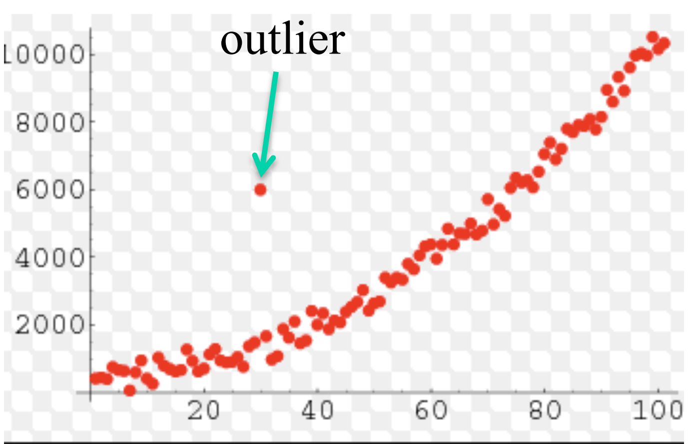
Salary = -50,000

Credit card transaction worth €3000

Temperature = 0 degrees and heating=off

Is an outlier an error?





Data quality issues

- Missing data – no attribute value
  - e.g., occupation=“ ”

Is missing data an error?

- Inconsistent data: containing discrepancies in codes or names, for example
  - Age=“42”; Birthday=“03/07/1997”
  - Different addresses for the same person

Data Characteristics

- **Pollution**
  - Occurs when a field starts to be used for a purpose other than what it was originally designed for, e.g. gender=C
  - Or when garbage gets into a data field
    - e.g. If data is transmitted from one data source to another, and errors in the data cause the information to be placed in the wrong attribute, e.g. . . .

Pollution

Name	Address	Date of Birth	Leaving Cert Y/N	Year

Mary Murphy, Dublin 15, 01/12/78, Y, 1996

Mary Murphy	Dublin 15	01/12/78	Y	1996

Mary Murphy, Dublin, 15, 01/12/78, Y, 1996

Mary Murphy	Dublin	15	01/12/78	Y

Summary of points covered to far . . .

- Before any analysis can be done on a dataset it is important to understand the characteristics of both the dataset and each individual attribute.
- Data understanding – how to collect the data, analyse the characteristics of the data set and analysis of the characteristics of each attribute in the data set
- Output from Data Understanding phase – report on the quality of the dataset.

Some of the terms covered so far:

Categorical

Ratio

Continuous

Discrete

Ordinal

Outliers

Interval

Inconsistent

Noise

Dimensionality

Nominal

Looking at Data using Rapidminer

- Your labsheets will explain how to read data into Rapidminer. The following slides will give you an idea of what to expect in the labs.
- The dataset being used here is a **Churn** (attrition) dataset from telecoms used to indicate a customer leaving the service of one company in favour of another company
- The dataset contains 21 fields worth of information, and about 3333 rows of data (3333 customers) along with an indication of whether or not that customer churned (left the company)

To begin with
Take a look at the field values for some of the records

Data view displays the data itself

Meta Data View · Data View · Plot View · Annotations

ExampleSet (3333 examples, 1 special attribute, 20 regular attributes)   View Filter (3333 / 3333): all

Row No.	Churn?	State	Account Le...	Area Code	Phone	Int'l Plan	VMail Plan	VMail Mes...	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge
1	False.	KS	128	415	382-4657	no	yes	25	265.100	110	45.070	197.400	99	16.780
2	False.	OH	107	415	371-7191	no	yes	26	161.600	123	27.470	195.500	103	16.620
3	False.	NJ	137	415	358-1921	no	no	0	243.400	114	41.380	121.200	110	10.300
4	False.	OH	84	408	375-9999	yes	no	0	299.400	71	50.900	61.900	88	5.260
5	False.	OK	75	415	330-6626	yes	no	0	166.700	113	28.340	148.300	122	12.610
6	False.	AL	118	510	391-8027	yes	no	0	223.400	98	37.980	220.600	101	18.750
7	False.	MA	121	510	355-9993	no	yes	24	218.200	88	37.090	348.500	108	29.620
8	False.	MO	147	415	329-9001	yes	no	0	157	79	26.690	103.100	94	8.760
9	False.	LA	117	408	335-4719	no	no	0	184.500	97	31.370	351.600	80	29.890
10	False.	WV	141	415	330-8173	yes	yes	37	258.600	84	43.960	222	111	18.870
11	True.	IN	65	415	329-6603	no	no	0	129.100	137	21.950	228.500	83	19.420

Row number

Attributes about the customer

Label: the column the model will try to predict

Next . . .
Take a look at information about each attribute

MetaData view displays information about each attribute

Meta Data View | Data View | Plot View | Annotations

ExampleSet (3333 examples, 1 special attribute, 20 regular attributes)

Role	Name	Type	Statistics	Range	Missings
label	Churn?	binominal	mode = False. (2850), least = True. (483)	False. (2850), True. (483)	0
regular	State	polynominal	mode = WV (106), least = CA (34)	WV (106), MN (84), NY (83), AL (80	0
regular	Account Length	integer	avg = 101.065 +/- 39.822	[1.000 ; 243.000]	0
regular	Area Code	integer	avg = 437.182 +/- 42.371	[408.000 ; 510.000]	0
regular	Account	polynominal	mode = 382-4657 (1), least = 382-4657 (1)	327-1058 (1), 327-1319 (1), 327-	0
regular	Int'l Plan	binominal	mode = no (3010), least = yes (323)	no (3010), yes (323)	0
regular	VMail Plan	binominal	mode = no (2411), least = yes (922)	yes (922), no (2411)	0
regular	VMail Message	integer	avg = 8.099 +/- 13.688	[0.000 ; 51.000]	0
regular	Day Mins	real	avg = 179.775 +/- 54.467	[0.000 ; 350.800]	0
regular	Day Calls	integer	avg = 100.436 +/- 20.069	[0.000 ; 165.000]	0
regular	Day Charge	real	avg = 30.562 +/- 9.259	[0.000 ; 59.640]	0
regular	Eve Mins	real	avg = 200.980 +/- 50.714	[0.000 ; 363.700]	0
regular	Eve Calls	integer	avg = 100.114 +/- 19.923	[0.000 ; 170.000]	0

Attribute name

Attribute data type

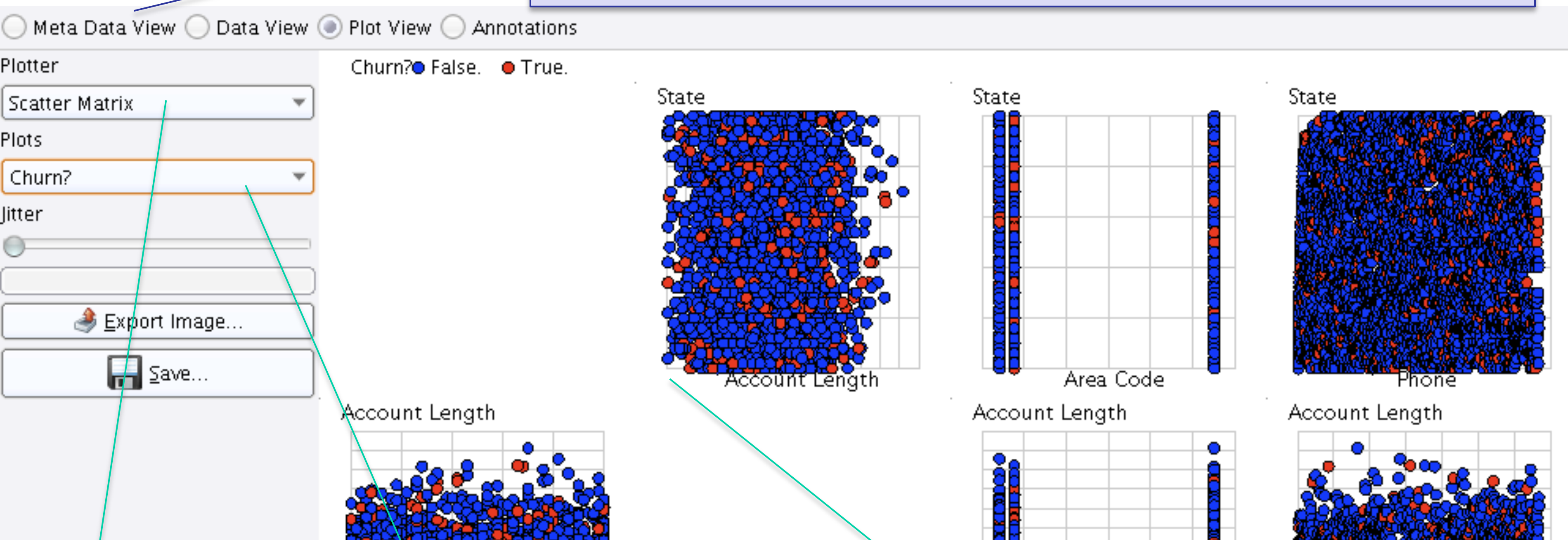
Attribute statistics

Number of missing values

Min and max values for numeric, list of values for alpha

The final view is. . .
A range of graphs you can plot from the data

Plot view: a range of graphs available to plot the data



A selection of types of graphs

Select the attributes to plot

Graph window

Types of graphs will be covered in the following slides

Initial data survey

Also called

Exploratory Data Analysis (EDA)

Covering . . . .

1. Basic statistics

2. Data visualisation (plots)

Exploratory Data Analysis

- Can carry out a survey of data to assess if there are patterns of interest in the data set
  - Done using graphing (data visualisation) and statistics
    - Simple (and not so simple) **graphs**, plots or tables are important tools in exploring important relationships in data
    - **Statistics** can analyse attribute properties and relationships between attributes

Descriptive Statistics and Plots

Descriptive Statistics:

- Mean or Mode
- Standard deviation
- Min & Max value
- Percentiles

Useful plots:

- Histograms and box plot (quartile plot) to look at the distribution of terms in one attribute
- Scatter plot to look at relationships between attributes
- Parallel plot to view all attribute values ranges

Example using RapidMiner

- The following slides use exploratory methods to delve into a *churn* data set. The attributes are as follows:

- State – categorical
- 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
- VoiceMail Plan - dichotomous categorical, yes or no
- Number of voice mail messages – integer valued
- Total day minutes – continuous, minutes customer used services during day

- Total day calls – integer valued
- Total day charge – continuous, perhaps based on foregoing two variables
- Total evening minutes – continuous, total minutes customer used the service during the evening
- Total evening calls – integer valued
- Total evening charge - continuous, perhaps based on foregoing two variables
- Total night minutes - continuous, total minutes customer used the service during the night
- Total night calls – integer valued
- Total night charge - continuous, perhaps based on foregoing 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 foregoing two variables
- Number of calls to customer services – integer valued

Exploration using statistics

- Summary statistics are usually calculated automatically by data mining tools when the data is read in, including:
- Numeric data: :
  - **Mean** – average value
  - **Standard deviation** – how far from the mean, on average, values lie
  - **Spread: min** value and **max** value (allows you to identify if outliers are distorting the mean and standard deviation).
- Nominal data:
  - **Mode**: most frequently occurring term
  - Value **frequencies** – the percentage of time that value occurs in the dataset.

Examples

Day Charge	real	avg = 30.562 +/- 9.259	[0.000 ; 59.640]

Values range from 0 to 59, mean is 30 with a standard deviation of 9 indicating a normal distribution around the mean, which is approximately the middle term.

VMail Message	integer	avg = 8.099 +/- 13.688	[0.000 ; 51.000]

Mean is only 8, but values range up to 51 indicating either skewed data or outlier values

State	polynominal	mode = WV (106), least = CA (34)	WV (106), MN (84), NY (83), AL (80), OH (78), OR (78), W

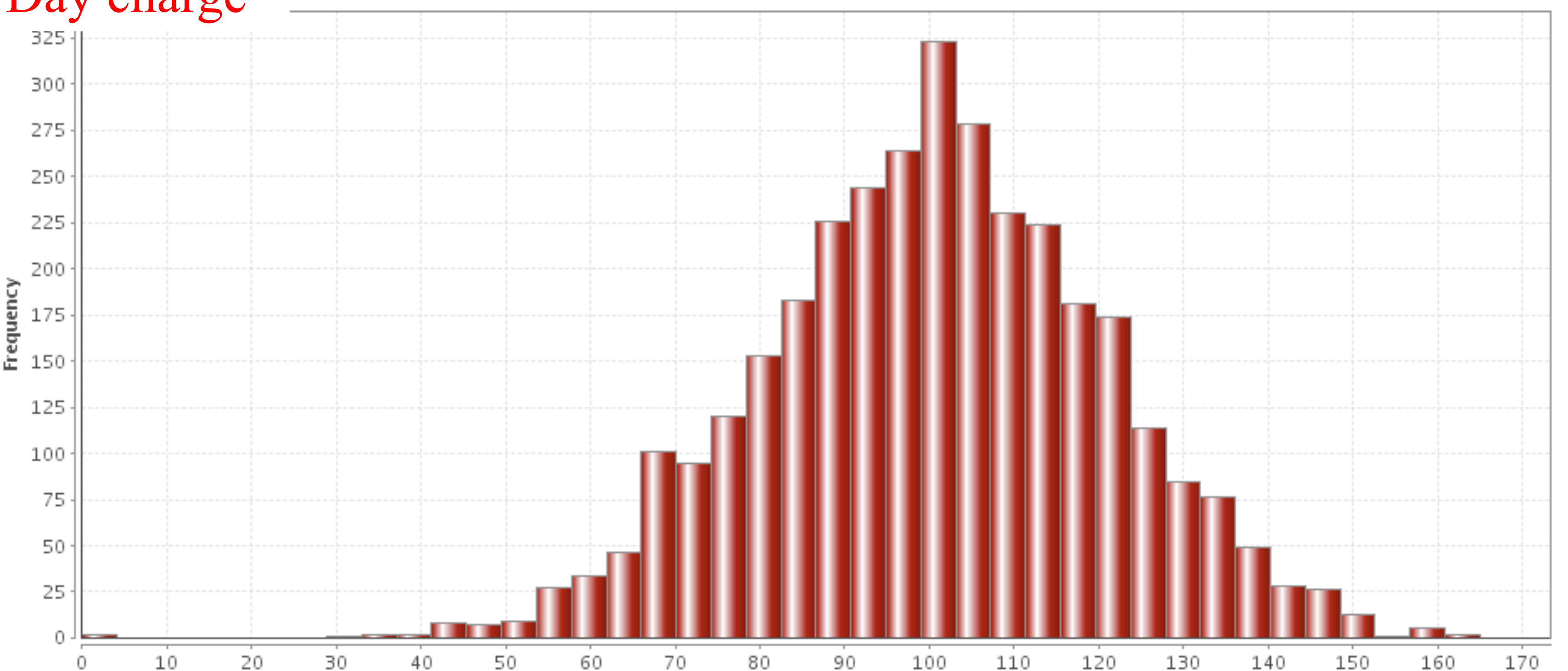
Categorical data showing mode, least frequent term, and term occurences (US states).

For all the statistics above, ideally a domain expert would evaluate if the distribution of values is what you would expect is such a dataset, or if more data is needed to give a true representation of the domain.

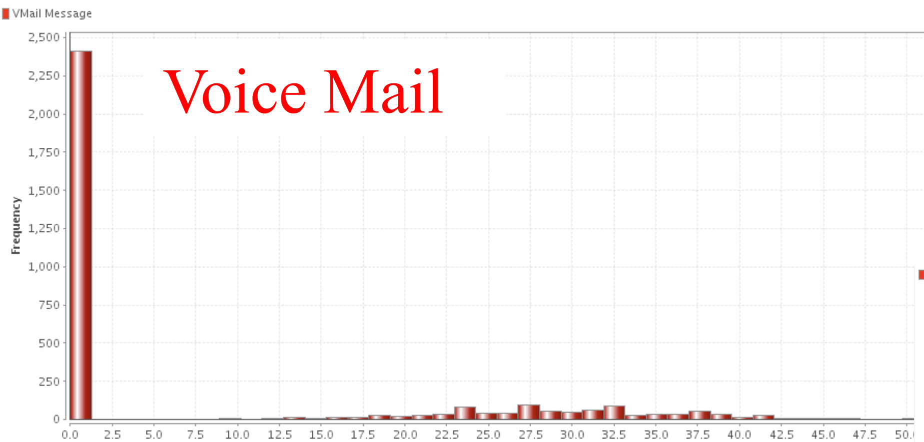
Histograms – data distribution

You can plot histograms of all columns to see if the range of values for that column is as you would expect. Ideally columns will have a normal distribution like this:

Day charge



Put sometimes they can be skewed left or right like this . . .



A domain expert would need to determine if the distribution made sense for this column, or if additional data is needed.

Percentiles – data distribution

- For continuous data, the notion of a percentile or quartiles is a useful evaluation of the distribution.

A **percentile** is a measure that tells us what percent of the numbers scored at or below that measure.

- For instance, the 50th percentile is the value such that 50% of all values are less than it, and 50% of all numbers are greater than it.

(Revision on Percentiles)

1, 3, 4, 6, 9, 10, 13, 14, 14, 16, 17, 18, 20, 21, 22, 26, 28, 30, 33, 33

Above is a list of 20 numbers.

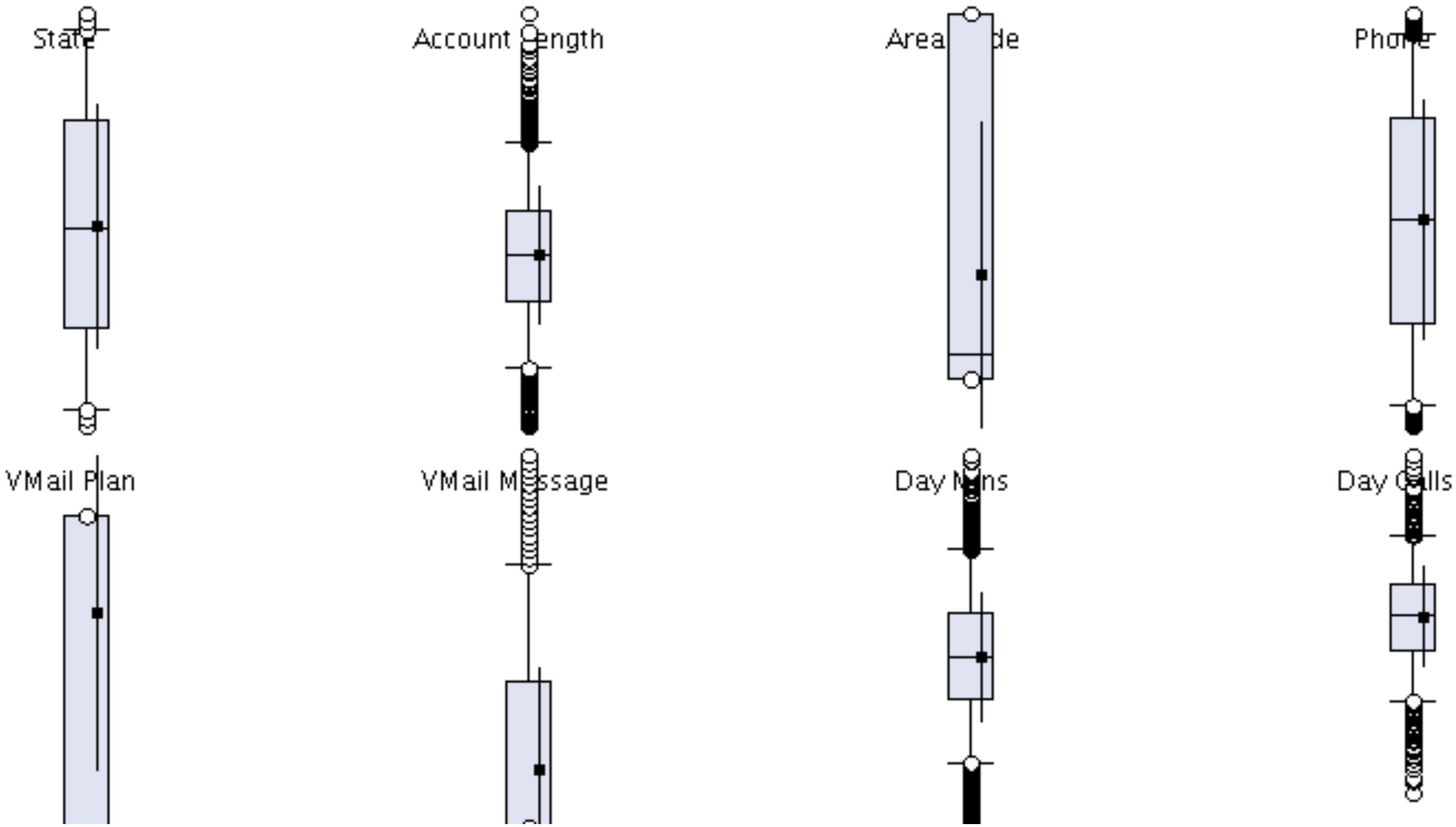
The 50th percentile is 16.5, as 50% of the numbers are lower than this, and 50% of the numbers are higher than this.

The 90th percentile is 30.3: 90% of numbers are lower than this number, and 10% of numbers are higher than it.

The 10th percentile is 3.9, as 10% of the numbers are lower than this number, and 90% are higher.

Other ways to represent this information . .

- A Quartile Color Matrix gives box plots for each attribute.



What does this mean . . .

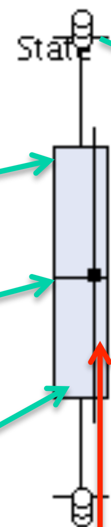
Explaining the box plot (called quartile in RapidMiner)

A box plot displays a lot of information about the distribution of data in an attribute, and the possibility of outliers (nicely explained on wikipedia).

25^{th} percentile

50^{th} percentile

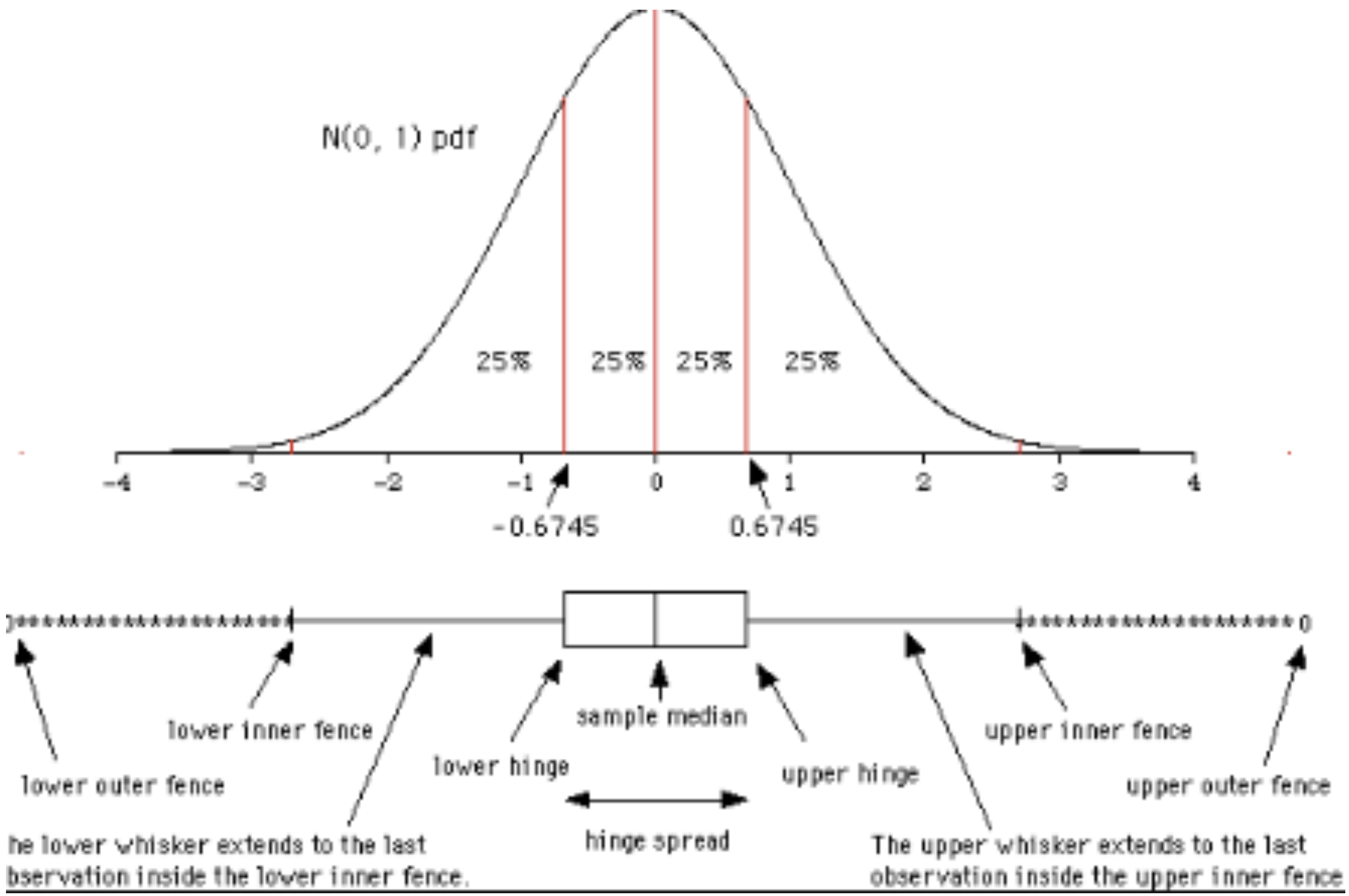
75^{th} percentile



The whiskers at each end (horizontal bar) represent values within normal range of the average, given the interquartile range. Anything outside of these (the circles) represent possible outlier values. They often represent the 10^{th} and 90^{th} percentile.

The vertical line and circle gives an indication of how close this distribution is to a normal distribution.

Box plots V histogram



Working with the Class label . . .

Remember: Overall objective for churn dataset: *to develop a model of the type of customer likely to churn*.

The class label is the attribute that distinguishes customers who have versus have not churned.

This dataset has examples of two classes (groups) of people:

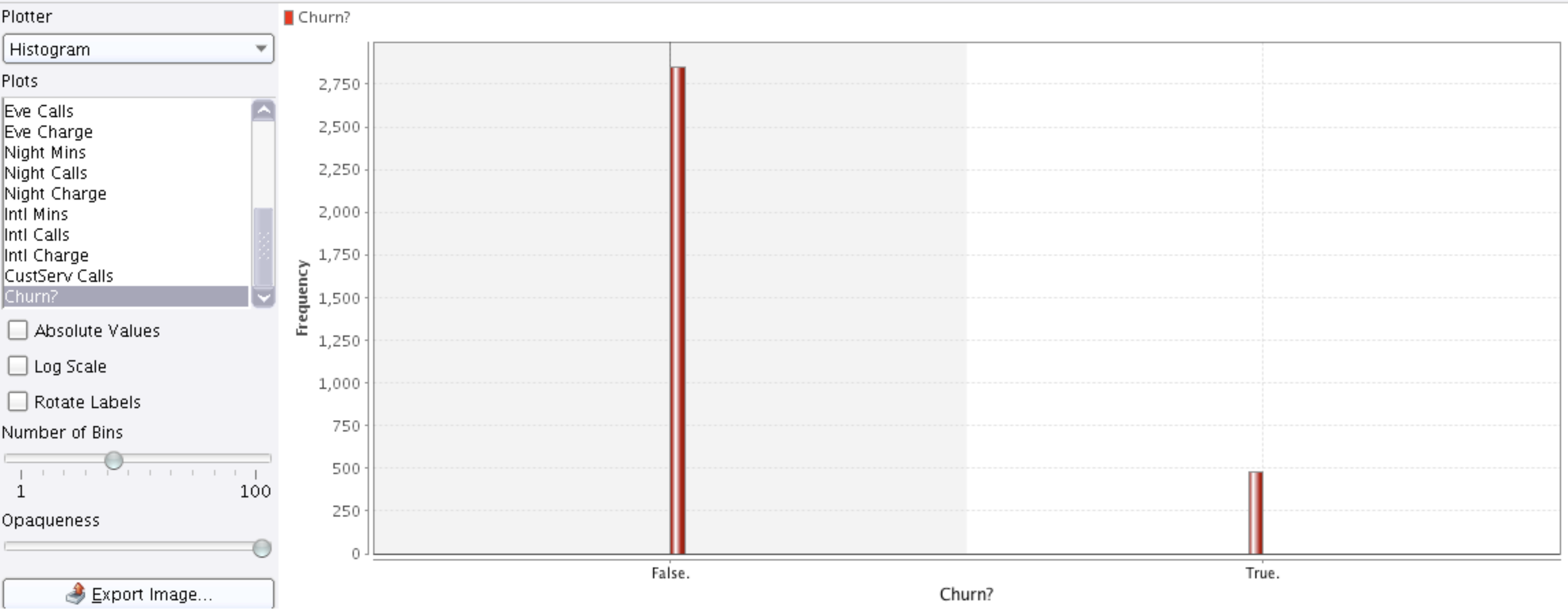
Those who have churned (churn = true)

and those who have not churned (churn=false)

An important step in EDA is to see if there are enough examples in the dataset of each class of customer?

Sufficient data?

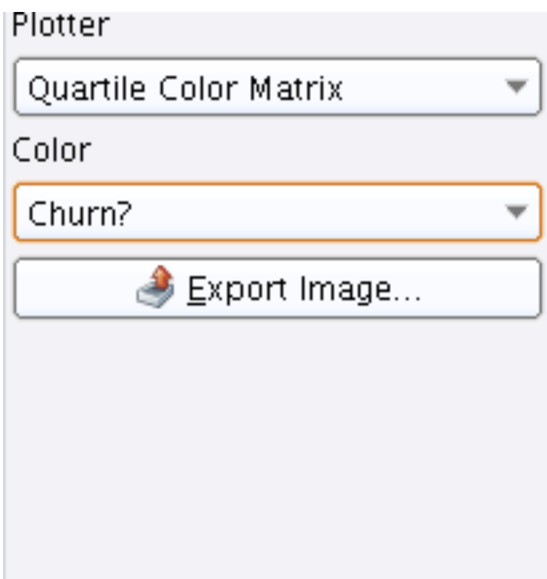
- Do see if there are enough examples of each class, we plot a HISTOGRAM of the attribute 'churn?', which is the class label

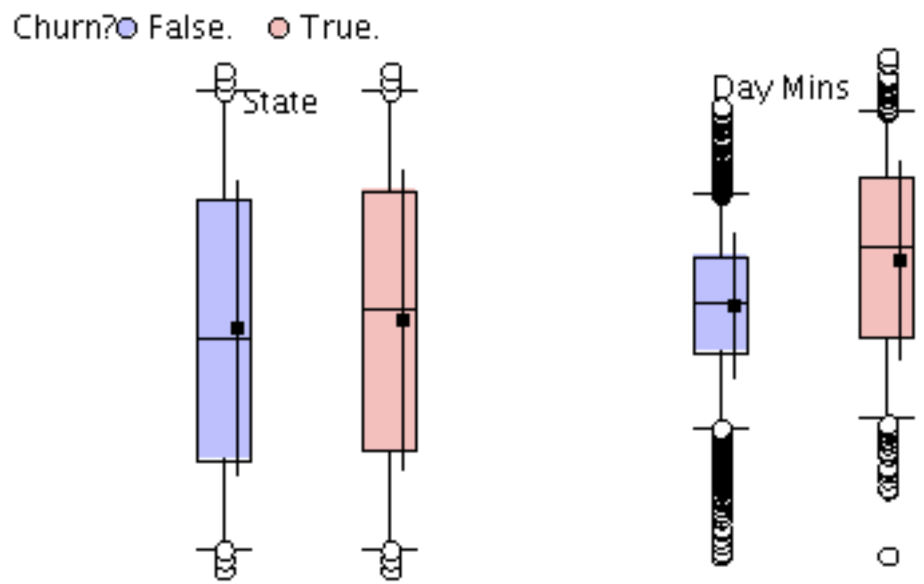


This dataset is very UNBALANCED which will cause problems . . .
A model could predict everyone as FALSE and be correct for most of the rows in the dataset.

Using box plots to identify predictive attributes

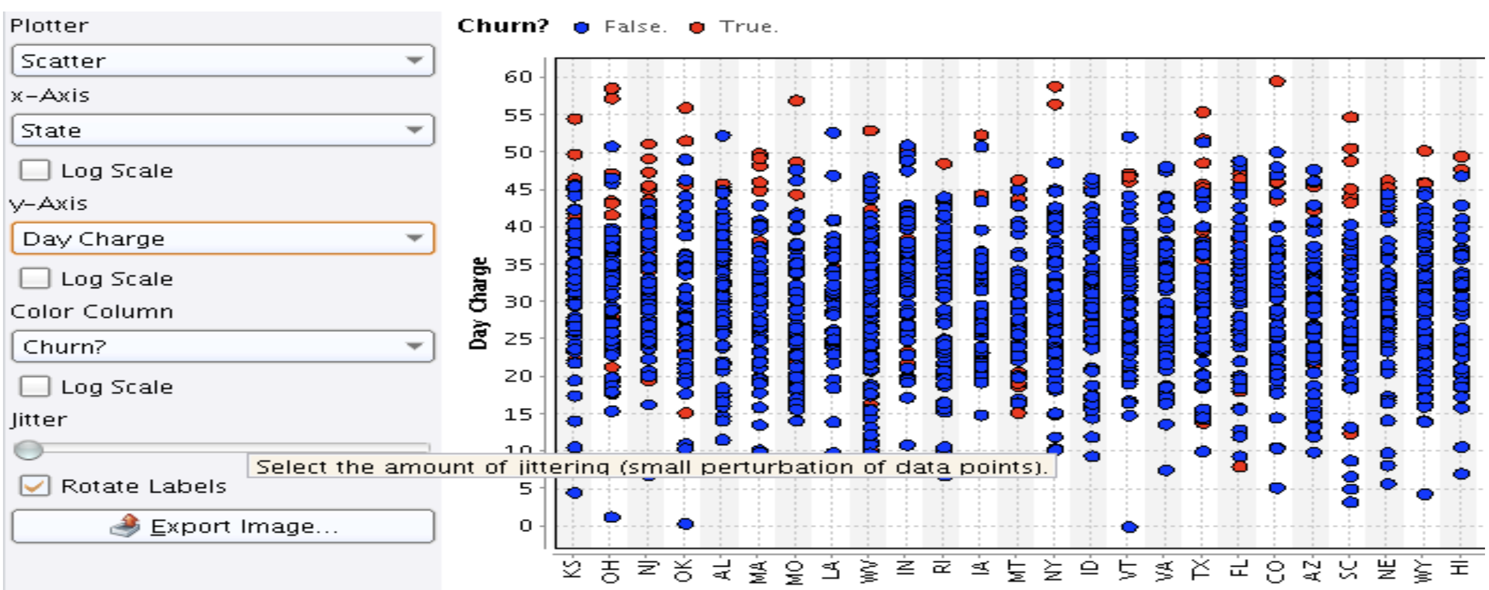
- On the left hand side, if you select the churn? Attribute under 'color', Rapidminer will generate separate box plots for churn = false data, and churn = true data. Where the box plots differ, it highlights an attribute that will be good at distinguishing between the two classes for customers



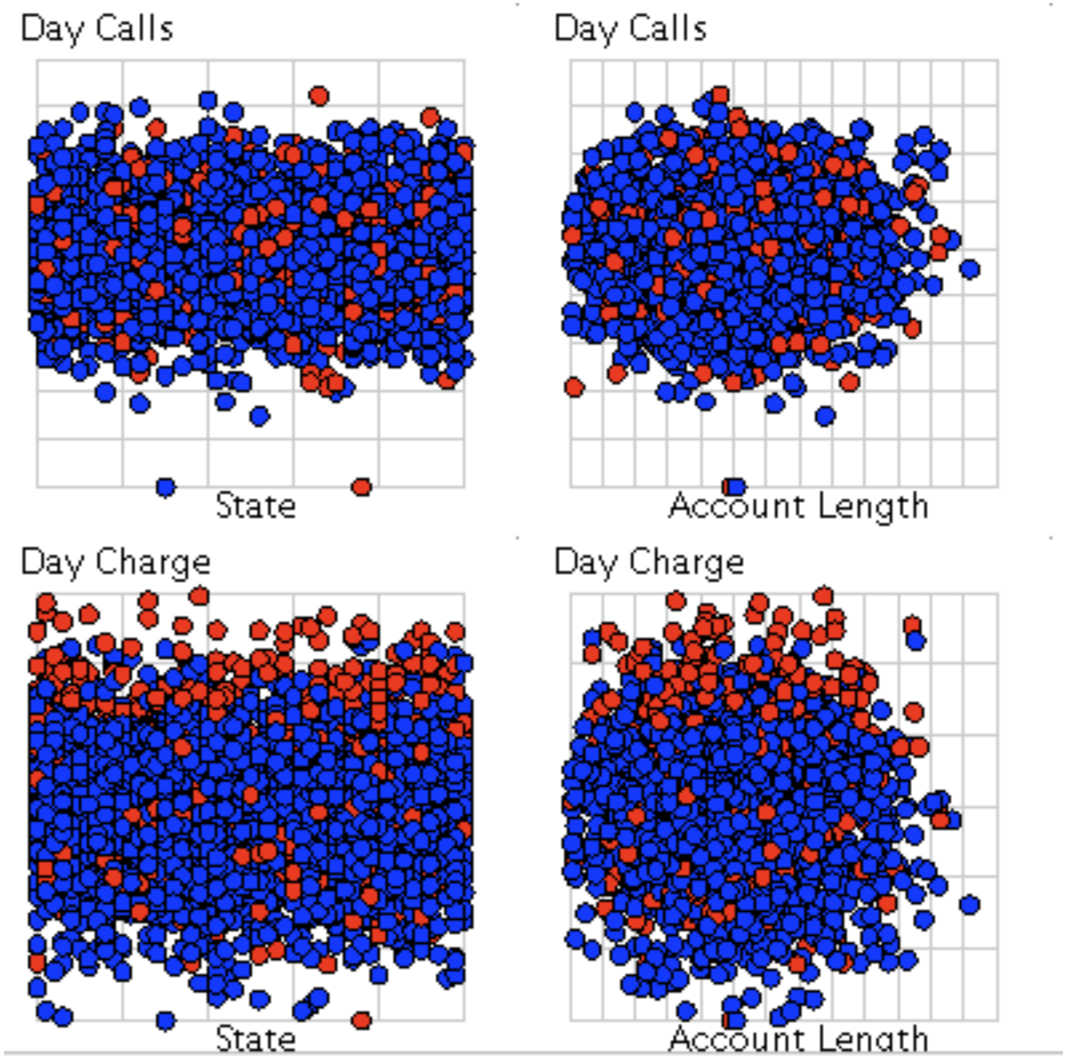


Scatter plots

- Another useful tool in identifying useful combinations of attributes is a scatter plot.
- A scatter plot plots one variable against another. The plot below plots day charges against US states, with color distinguishing between customers that churned or stayed with company. In this kind of plot, you are looking for areas of just red or just blue. In this case, high day charge in all states represent customers that churned.

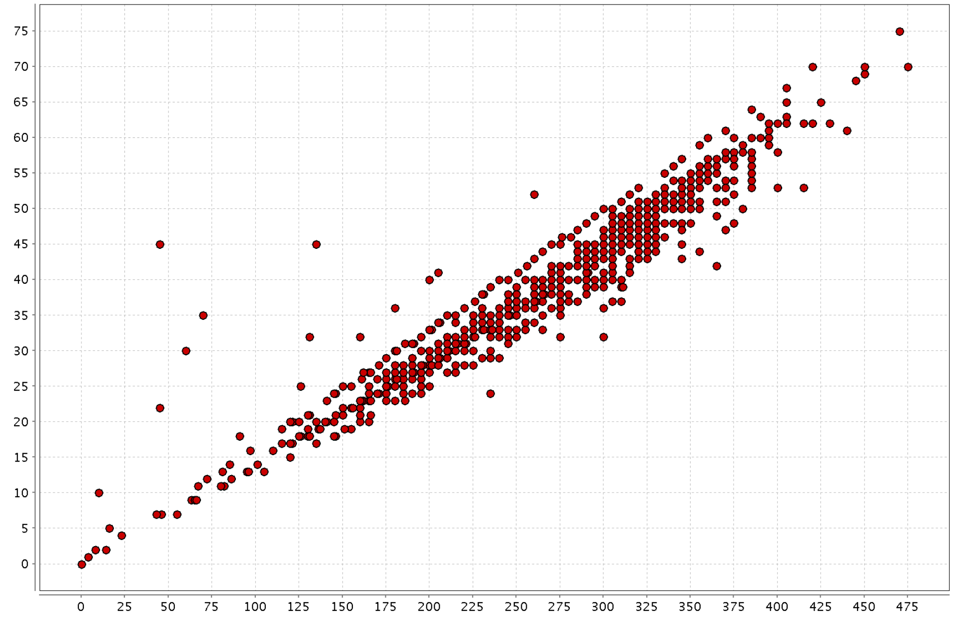


A scatter matrix shows all possible scatter plots in the data set. It takes a while to generate . . .



Correlated attributes

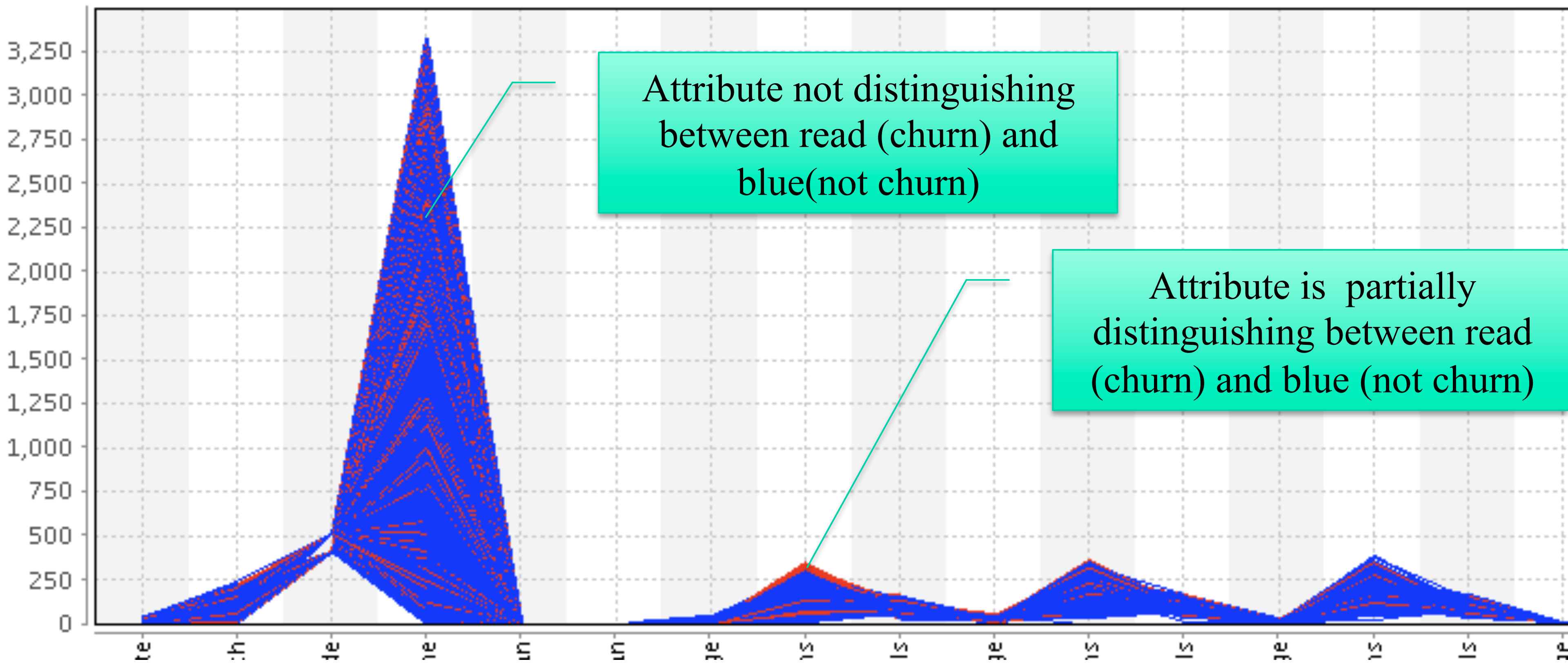
- A scatter plot that is a diagonal line indicates two attributes that are highly correlation, i.e. the hold the same information.
  - If one has a high value so does the other, and vice versa.
- Where two attributes have a high correlation, one can be deleted from the dataset.



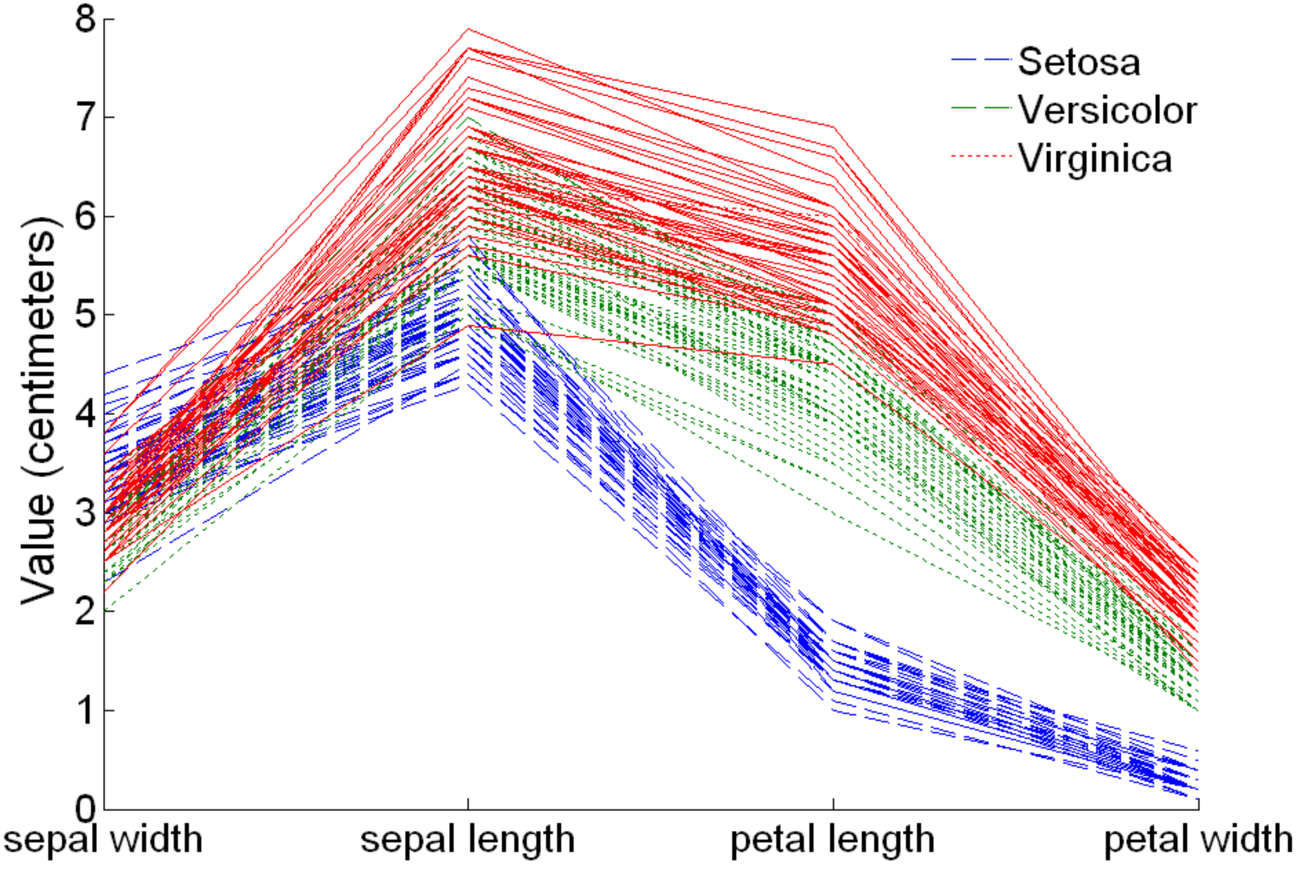
Parallel plot

- Used to plot the attribute values of high-dimensional data
- Uses a set of parallel axes
- The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
- Thus, **each object is represented as a line**
- Often, the lines representing a distinct class of object group together, at least for some attributes
- Ordering of attributes can be important in seeing such groupings

Parallel plot for churn dataset:



Parallel plot for iris dataset – distinguishing between three different types of iris based on petal and sepal size



Other visualisation techniques

- Rapid Miner has a selection of other graphs (plots) many of which are just alternative ways of showing the same information.
- Have a look at more of them in the lab. . .

Summary: Exploratory Data Analysis

- Exploratory data analysis will give you this first indication of:
  - If you have sufficient data
    - Enough data to represent each class
    - Enough data to correctly represent the range of values for each variable
  - If the attributes you have will be useful in predicting the class variable