Data Visualization

การสำรวจและนำเสนอข้อมูลหลายมิติสำหรับงาน Machine Learning

Basic Charts, Distribution, Multidimensional Visualization

บทนำ

คำกล่าวที่ว่า "ภาพหนึ่งภาพมีค่าเท่ากับคำพูดพันคำ" หมายถึงความสามารถในการบีบอัดข้อมูลที่กระจัดกระจายให้อยู่ในรูปแบบกราฟิกที่กระชับและเข้าใจได้อย่างรวดเร็ว

ในกรณีของตัวเลข Data Visualization และการสรุปเชิงตัวเลขช่วยให้เรามีเครื่องมือที่ทรงพลังในการสำรวจข้อมูลและนำเสนอผลลัพธ์อย่างมีประสิทธิภาพ

บทบาทของ Visualization ใน Machine Learning

เทคนิค Visualization ถูกใช้เป็นหลักในส่วน Preprocessing ของกระบวนการ Machine Learning โดยมีบทบาทสำคัญดังนี้:

Data Cleaning

ค้นหาค่าที่ไม่ถูกต้อง ค่าที่หายไป แถวที่ซ้ำกัน และคอลัมน์ที่มีค่าเดียวกันทั้งหมด

Attribute Derivation

ช่วยกำหนดว่า Attributes ใดควรรวมอยู่ในการวิเคราะห์และ Attributes ใดที่อาจซ้ำซ้อน

Binning

ช่วยกำหนดขนาด Bin ที่เหมาะสมหาก Numerical Attributes ต้องการการแบ่งกลุ่ม

วัตถุประสงค์ของบทนี้

เรามุ่งเน้นการใช้การนำเสนอแบบกราฟิกเพื่อวัตถุประสงค์ของ **Data Exploration** โดยเฉพาะอย่างยิ่งที่เกี่ยวข้องกับ Predictive Analytics

แม้ว่าจุดสนใจของเราจะไม่ได้อยู่ที่ Visualization เพื่อการรายงานข้อมูล แต่บทนี้ก็นำเสนอแนวคิดเกี่ยวกับประสิทธิผลของการแสดงผลแบบกราฟิกต่างๆ เพื่อวัตถุประสงค์ในการนำเสนอข้อมูล

ความหมายของคำว่า "Graph"

หมายเหตุ: คำว่า "graph" มีความหมายสองแบบในทาง statistics

- หมายถึงรูปแบบต่างๆ ในการแสดงข้อมูล (เช่น Line Chart, Bar Plot, Histogram)
- หมายถึง Data Structure และ Visualization ใน Networks

การใช้คำว่า "Chart" หรือ "Plot" สำหรับ Visualizations ที่เราสำรวจในบทนี้จะช่วยหลีกเลี่ยงความสับสน

Data Exploration: ขั้นตอนที่จำเป็น

Data Exploration เป็นขั้นตอนเริ่มต้นที่จำเป็นไม่ว่าจะมีการวิเคราะห์ที่เป็นทางการมากขึ้นตามมาหรือไม่

Free-form Exploration

- เข้าใจโครงสร้างข้อมูล
- ทำความสะอาดข้อมูล
- ระบุ Outliers
- ค้นพบ Patterns เบื้องต้น
- สร้างคำถามที่น่าสนใจ

Focused Exploration

- มุ่งเน้นไปที่คำถามเฉพาะที่สนใจ
- สนับสนุนเป้าหมายที่เฉพาะเจาะจง
- ใช้ร่วมกับ Free-form Exploration

มุมมองตาม Machine Learning Goal

การอภิปรายของเราจะมาจากมุมมองของว่า Visualization สนับสนุนเป้าหมาย Machine Learning ที่ตามมาอย่างไร โดยเฉพาะอย่างยิ่ง เราแยกความแตกต่างระหว่าง:

Supervised Learning

Classification (Categorical Target) และ Prediction (Numerical Target)

Unsupervised Learning

Data Reduction และ Clustering

ตัวอย่างที่ 1: Boston Housing Data

Boston Housing Data ประกอบด้วยข้อมูลเกี่ยวกับ Census Tracts ในบอสตันซึ่งมีการวัดหลายอย่าง (เช่น อัตราอาชญากรรม อัตราส่วนนักเรียนต่อครู)

มี 14 Attributes รวมถึง Attribute ที่สร้างขึ้นใหม่ชื่อ CAT.MEDV ซึ่งถูกสร้างโดยการจัดหมวดหมู่ Median Value (MEDV) เป็นสองหมวดหมู่: สูงและต่ำ

Attributes ใน Boston Housing Dataset

TABLE 3.1 DESCRIPTION OF ATTRIBUTES IN BOSTON HOUSING DATASET

CRIM	Crime rate
ZN	Percentage of residential land zoned for lots over 25,000 ft^2
INDUS	Percentage of land occupied by nonretail business
CHAS	Does tract bound Charles River (= 1 if tract bounds river, = 0 otherwise)
NOX	Nitric oxide concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Percentage of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property tax rate per $10,000
PTRATIO	Pupil-to-teacher ratio by town
LSTAT	Percentage of lower status of the population
MEDV	Median value of owner-occupied homes in $1000s
CAT.MEDV	Is median value of owner-occupied homes in tract above $30,000 (CAT.MEDV = 1) or not (CAT.MEDV = 0)

Row No.	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV	CAT. MEDV
1	0.006	18	2.310	0	0.538	6.575	65.200	4.090	1	296	15.300	4.980	24	0
2	0.027	0	7.070	0	0.469	6.421	78.900	4.967	2	242	17.800	9.140	21.600	0
3	0.027	0	7.070	0	0.469	7.185	61.100	4.967	2	242	17.800	4.030	34.700	1
4	0.032	0	2.180	0	0.458	6.998	45.800	6.062	3	222	18.700	2.940	33.400	1
5	0.069	0	2.180	0	0.458	7.147	54.200	6.062	3	222	18.700	5.330	36.200	1
6	0.030	0	2.180	0	0.458	6.430	58.700	6.062	3	222	18.700	5.210	28.700	0
7	0.088	12.500	7.870	0	0.524	6.012	66.600	5.561	5	311	15.200	12.430	22.900	0
8	0.145	12.500	7.870	0	0.524	6.172	96.100	5.950	5	311	15.200	19.150	27.100	0
9	0.211	12.500	7.870	0	0.524	5.631	100	6.082	5	311	15.200	29.930	16.500	0

FIGURE 3.1 FIRST NINE RECORDS IN THE BOSTON HOUSING DATA

3 Tasks ที่เป็นไปได้กับ Boston Housing Data

01

Supervised Predictive Task

Target Attribute คือ Median Value ของบ้านใน Tract (MEDV)

02

Supervised Classification Task

Target Attribute คือ Binary Attribute CAT.MEDV ที่บอกว่าค่าบ้านอยู่เหนือหรือต่ำกว่า $30,000

03

Unsupervised Task

เป้าหมายคือการทำ Clustering ของ Census Tracts

หมายเหตุ: MEDV และ CAT.MEDV ไม่ถูกใช้ร่วมกันในทั้งสามกรณี

ตัวอย่างที่ 2: Ridership on Amtrak Trains

Amtrak บริษัทรถไฟของสหรัฐอเมริกา รวบรวมข้อมูลเกี่ยวกับจำนวนผู้โดยสารเป็นประจำ

ที่นี่เรามุ่งเน้นการพยากรณ์จำนวนผู้โดยสารในอนาคตโดยใช้ Series ของจำนวนผู้โดยสารรายเดือนระหว่างเดือนมกราคม 1991 ถึงมีนาคม 2004

ดังนั้น Task ของเราที่นี่คือ Numerical Time Series Forecasting

Basic Charts

Charts พื้นฐานที่มีประสิทธิภาพมากที่สุดสามแบบคือ Bar Charts, Line Charts และ Scatter Plots

Charts เหล่านี้ง่ายต่อการสร้างใน RapidMiner และเป็นที่นิยมใช้มากที่สุดในโลกธุรกิจปัจจุบัน ทั้งใน Data Exploration และการนำเสนอ

หมายเหตุ: แม้ว่า Pie Charts จะเป็นที่นิยม แต่โดยปกติแล้วเป็น Visualizations ที่ไม่มีประสิทธิภาพ

Support Data Exploration by Basic Charts

Basic Charts สนับสนุน Data Exploration โดยการแสดงหนึ่งหรือสอง Columns ของข้อมูล (Attributes) ในแต่ละครั้ง

สิ่งนี้มีประโยชน์ในขั้นตอนแรกของการทำความคุ้นเคยกับ:

- โครงสร้างข้อมูล
- จำนวนและประเภทของ Attributes
- ปริมาณและประเภทของ Missing Values

How to use Basic Charts in Machine Learning Task

Supervised Learning

จะมีการมุ่งเน้นมากขึ้นที่ Target Attribute

ใน Scatter Plots, Target Attribute มักจะเชื่อมโยงกับแกน *y*

Unsupervised Learning

สำหรับ Data Reduction หรือ Clustering

Basic Charts ที่ถ่ายทอด Relationships (เช่น Scatter Plots) เป็นที่ต้องการ

การสร้าง Charts ใน RapidMiner

ใน RapidMiner, Charts โดยทั่วไปถูกสร้างผ่าน Visualizations Tab ใน Results View

Visualizations Tab

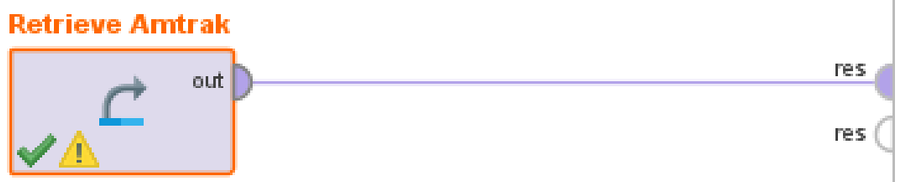
Interface หลักสำหรับสร้าง Charts ใน Results View

Turbo Prep View

Interface ที่สามารถเรียกใช้จาก Results View สำหรับ Data Preparation Tasks และ Charts Panel

Pivot Table

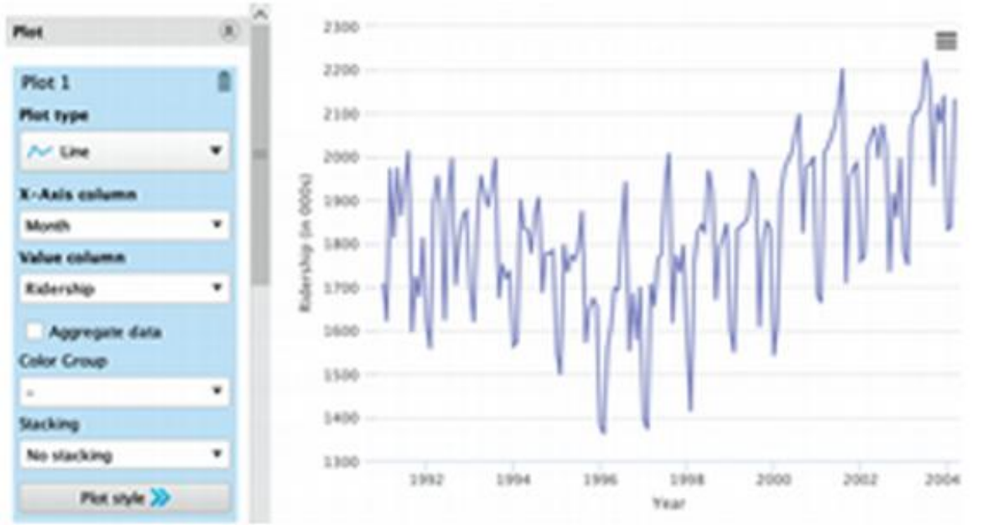
Turbo Prep View ให้วิธีที่รวดเร็วในการสร้าง Aggregation เช่น Pivot Table จากนั้น Visualize เป็น Table หรือ Bar Chart



Line Charts

Line Charts ใช้เป็นหลักสำหรับแสดง Time Series

การเลือก Time Frame ที่จะ Plot รวมถึง Temporal Scale ควรขึ้นอยู่กับ Horizon ของ Forecasting Task และลักษณะของข้อมูล



ตัวอย่างด้านบนแสดง Line Chart สำหรับ Time Series ของผู้โดยสารรถไฟรายเดือนบน Amtrak

Bar Charts: การเปรียบเทียบระหว่างกลุ่ม

Bar Charts มีประโยชน์สำหรับการเปรียบเทียบ Statistic เดียว (เช่น Average, Count, Percentage) ระหว่างกลุ่ม

ความสูงของ Bar (หรือความยาวในการแสดงแนวนอน) แสดงถึงค่าของ Statistic และ Bars ที่แตกต่างกันสอดคล้องกับกลุ่มที่แตกต่างกัน

สำคัญ: แกน x บน Bar Chart ต้องใช้เฉพาะสำหรับ Categorical Attributes เท่านั้น เพราะลำดับของ Bars ใน Bar Chart ควรสามารถสลับที่กันได้

Bar Charts: Numerical vs. Categorical Attributes

Numerical Attribute (MEDV)

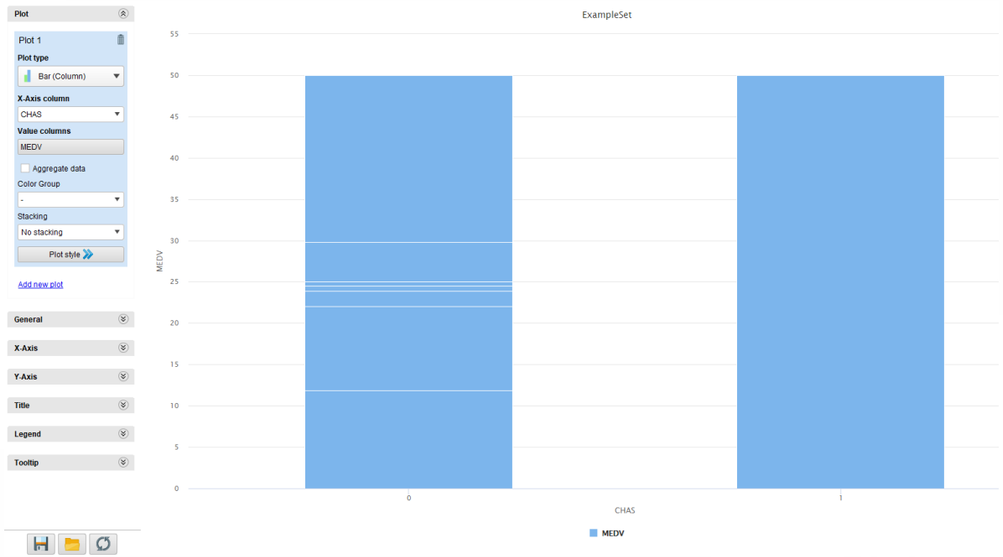
ใช้ Average MEDV บนแกน *y*

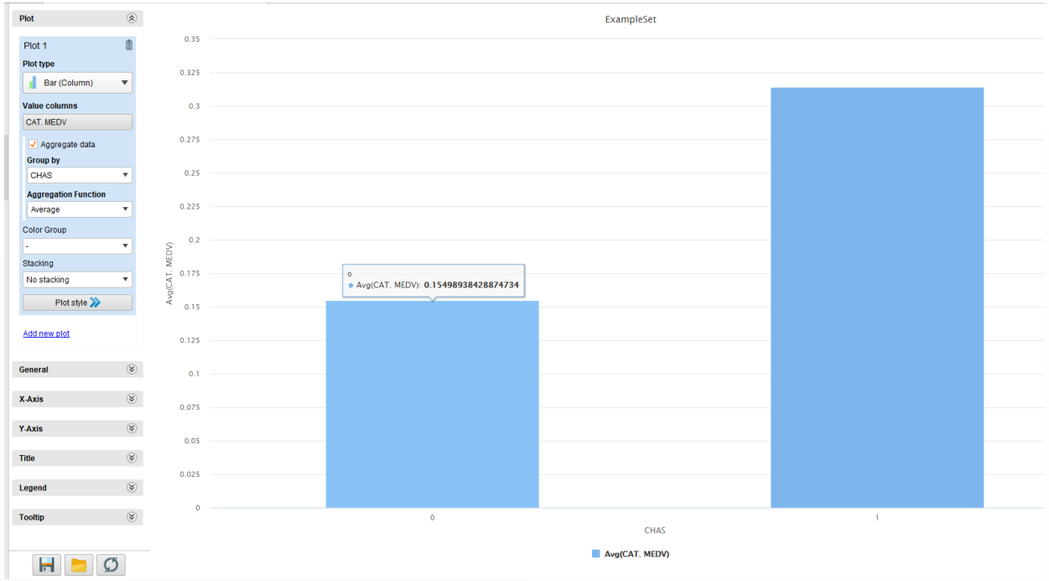
สนับสนุน Predictive Task: Numerical Target อยู่บนแกน *y* และแกน *x* ใช้สำหรับ Categorical Predictor ที่เป็นไปได้

Categorical Attribute (CAT.MEDV)

แกน *y* บอกเปอร์เซ็นต์ของ Tracts ที่มี MEDV เหนือ $30K

แกน *x* เป็น Binary Attribute ที่บอกความใกล้กับแม่น้ำ Charles



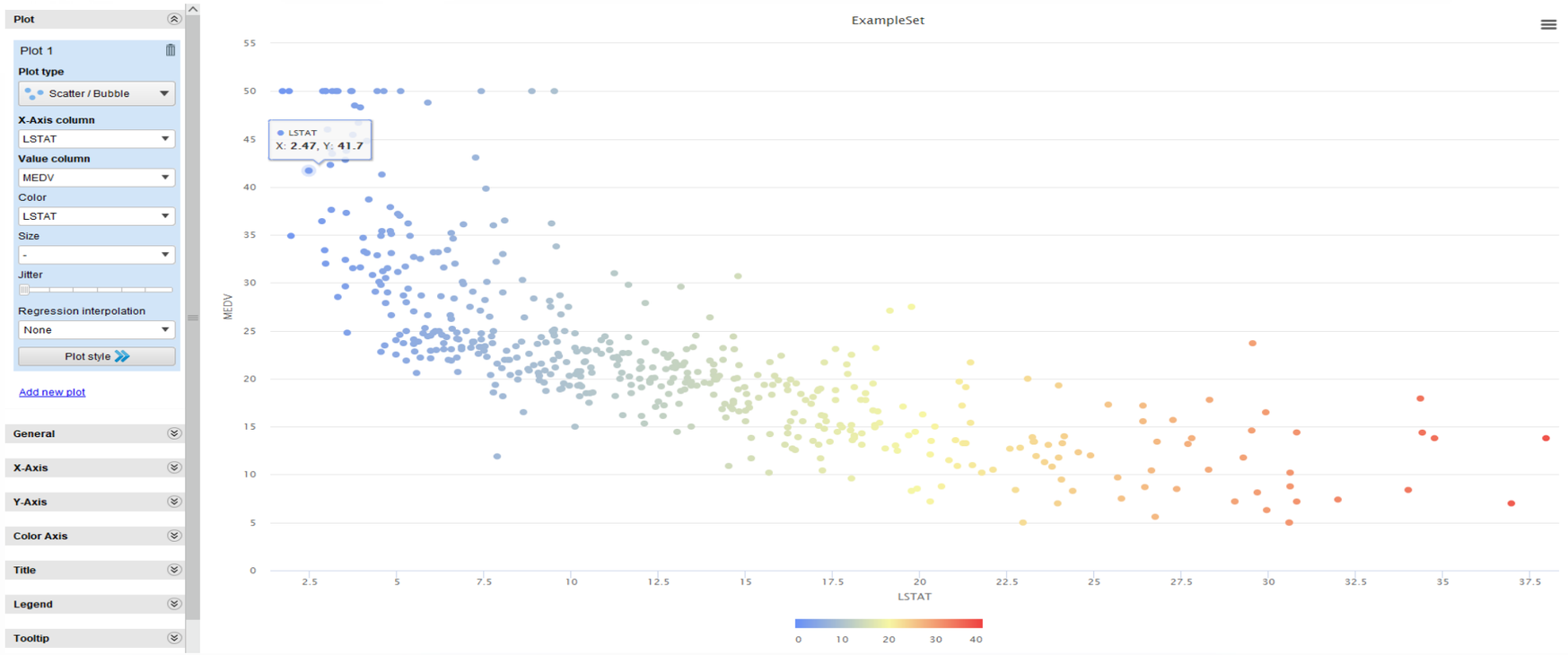


Scatter Plots

Scatter Plot แสดง MEDV vs. LSTAT ซึ่งเป็น Plot ที่สำคัญใน Prediction Task

โปรดสังเกตว่า Target MEDV อยู่บนแกน y และ LSTAT บนแกน x เป็น Predictor ที่เป็นไปได้

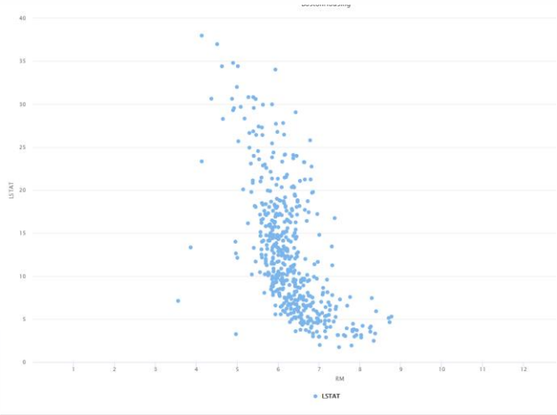
เนื่องจาก Attributes ทั้งสองใน Basic Scatter Plot ต้องเป็น Numerical จึงไม่สามารถใช้แสดง Relation ระหว่าง CAT.MEDV และ Predictors ที่เป็นไปได้สำหรับ Classification Task



การใช้ Scatter Plots ใน Unsupervised Learning

สำหรับ Unsupervised Learning, Scatter Plot ช่วยศึกษา Association ระหว่าง Numerical Attributes สองตัวในแง่ของ Information Overlap

นอกจากนี้ยังช่วยระบุ Clusters ของ Observations



Charts พื้นฐานทั้งสามแบบเน้นข้อมูล Global เช่น:

- ระดับโดยรวมของจำนวนผู้โดยสารหรือ MEDV
- การเปลี่ยนแปลงตามเวลา (Line Chart)
- ความแตกต่างระหว่าง Subgroups (Bar Chart)
- Relationships ระหว่าง Numerical Attributes (Scatter Plot)

ข้อจำกัดของ Basic Charts

จุดอ่อนหลักของ Basic Charts และ Distribution Plots ในรูปแบบพื้นฐาน (คือการใช้ตำแหน่งเทียบกับแกนเพื่อ Encode ค่า) คือพวกมันสามารถแสดงเพียงสอง Attributes เท่านั้น

ดังนั้นจึงไม่สามารถเปิดเผยข้อมูลหลายมิติได้

Basic Charts แต่ละแบบมีสองมิติ โดยแต่ละมิติถูกอุทิศให้กับ Attribute เดียว

Distribution Plots: Boxplots และ Histograms

ก่อนที่จะไปยัง Visualizations ที่ซับซ้อนกว่า เราต้องทราบ Charts สำคัญสองแบบที่แสดง Distribution ทั้งหมดของ Numerical Attribute

Boxplot

แสดง Distribution ทั้งหมดและมีประสิทธิภาพในการเปรียบเทียบ Subgroups

Histogram

แสดง Frequencies ของค่าทั้งหมดด้วย Vertical Connected Bars

ทำไมต้องใช้ Distribution Plots?

แม้ว่า Averages จะเป็นที่นิยมและมีประโยชน์ แต่โดยปกติแล้วมีประโยชน์มากที่จะดูที่ Statistics เพิ่มเติมเช่น Median และ Standard Deviation

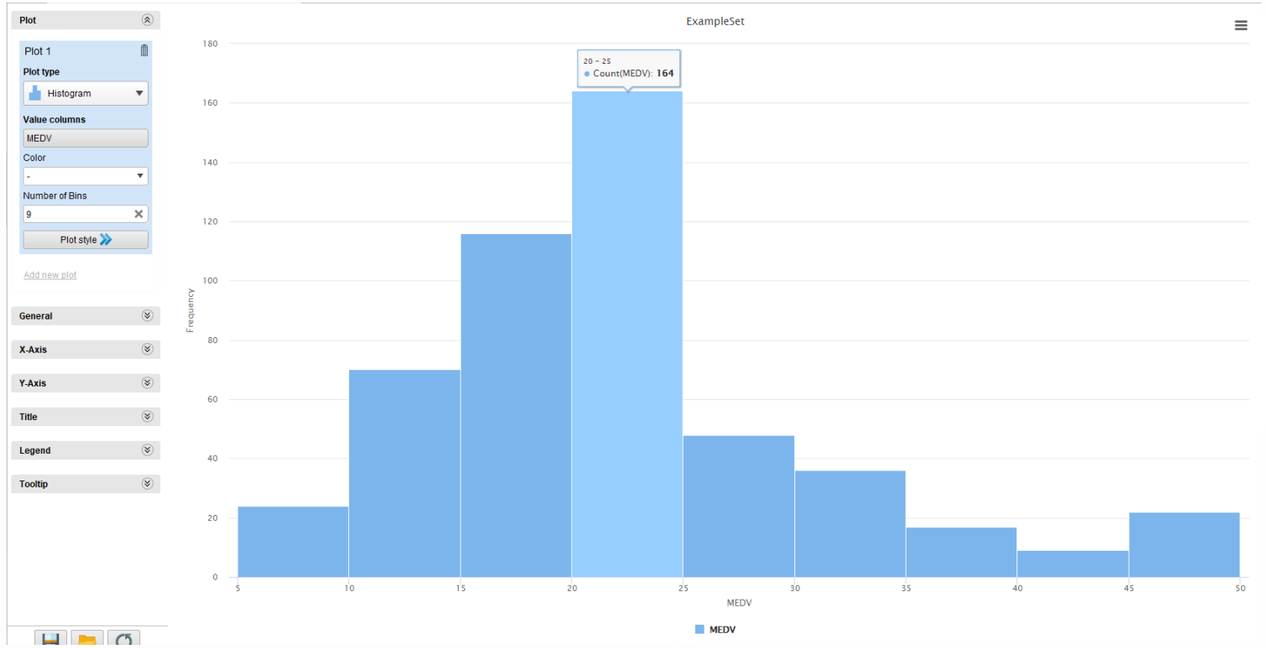
และยิ่งไปกว่านั้นคือการตรวจสอบ Distribution ทั้งหมด

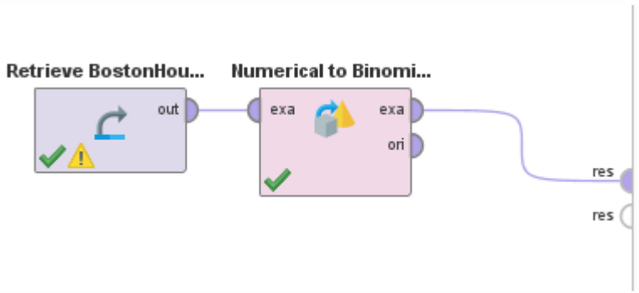
ข้อดี: ในขณะที่ Bar Charts สามารถใช้ Aggregation เดียวเท่านั้น Boxplots และ Histograms แสดง Distribution ทั้งหมดของ Numerical Attribute

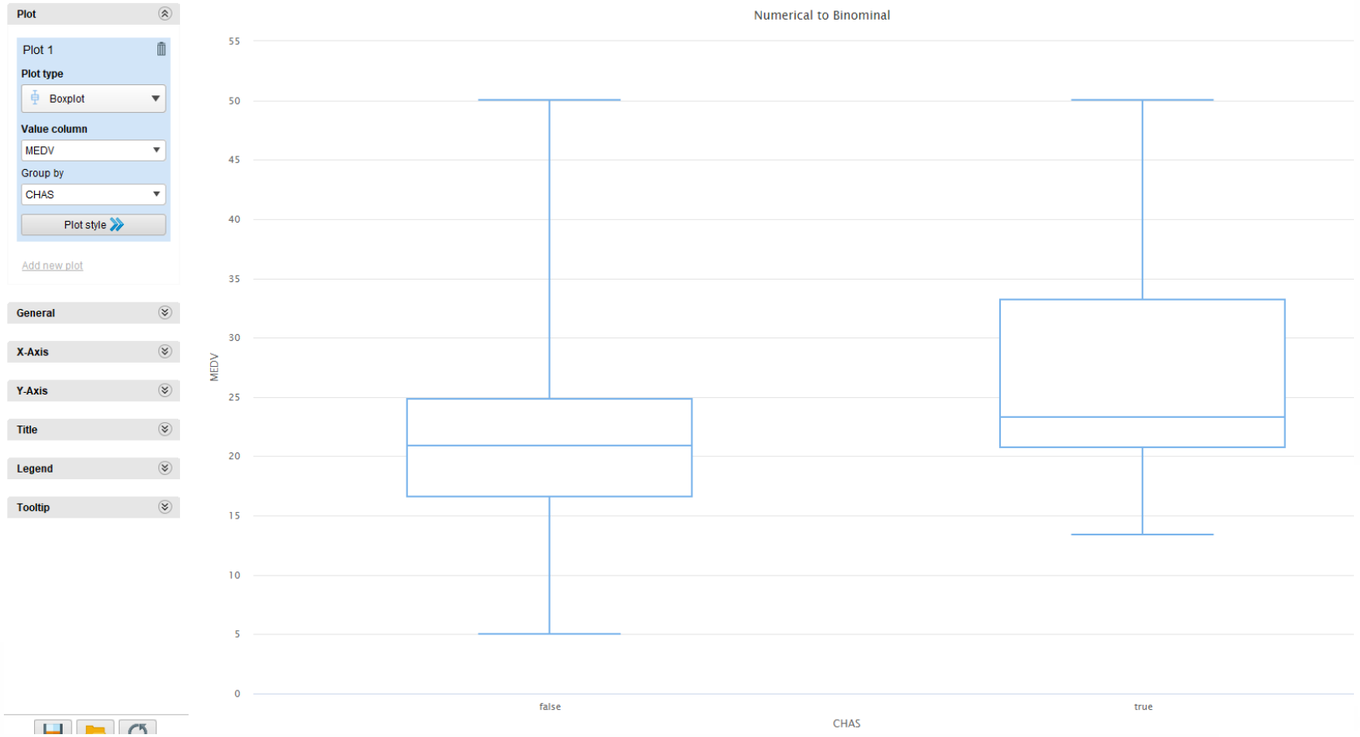
Histogram: ตัวอย่าง

Histogram แสดง Frequencies ของค่า *x* ทั้งหมดด้วย Series ของ Vertical Connected Bars

ตัวอย่างเช่น ในรูปด้านบน มี Tracts มากกว่า 150 แห่งที่ MEDV อยู่ระหว่าง $20K และ $25K







Boxplot: โครงสร้าง

Boxplot แสดง Attribute ที่กำลัง Plot บนแกน y (แม้ว่า Plot อาจหมุน 90° เพื่อให้ Boxes ขนานกับแกน x)

Box

ครอบคลุม 50% ของข้อมูล ตัวอย่างเช่น ใน Box ด้านขวา ครึ่งหนึ่งของ Tracts มี MEDV ระหว่าง $20,000 และ $33,000

Median Line

เส้นแนวนอนภายใน Box แสดง Median (50th Percentile)

Quartiles

ด้านบนและด้านล่างของ Box แสดง 75th และ 25th Percentiles ตามลำดับ

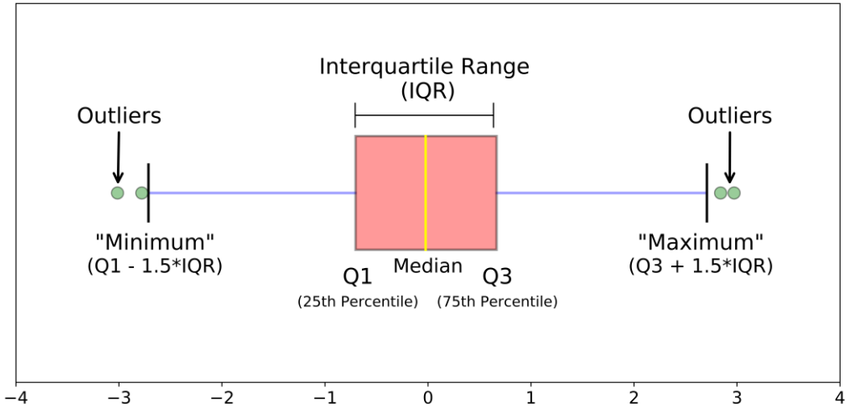
Boxplot: องค์ประกอบเพิ่มเติม

เส้นที่ขยายไปด้านบนและด้านล่างของ Box ครอบคลุม Data Range ที่เหลือ

Outliers อาจถูกแสดงเป็นจุดหรือวงกลม

บางครั้ง Average ถูกทำเครื่องหมายด้วยเครื่องหมาย + (หรือคล้ายกัน)

การเปรียบเทียบ Average และ Median ช่วยในการประเมินว่าข้อมูล Skewed แค่ไหน



การใช้ Distribution Plots ใน Prediction Tasks

เนื่องจาก Histograms และ Boxplots มุ่งเน้นไปที่ Numerical Attributes รูปแบบพื้นฐานของพวกมันจึงมีประโยชน์สำหรับ **Prediction Tasks**

Histogram ของ MEDV เผยให้เห็น Skewed Distribution

การแปลง Target Attribute เป็น log(MEDV) อาจปรับปรุงผลลัพธ์ของ Linear Regression Predictor

Boxplots สำหรับ Unsupervised Learning

Boxplots สามารถสนับสนุน **Unsupervised Learning** โดยการแสดง Relationships ระหว่าง Numerical Attribute (แกน y) และ Categorical Attribute (แกน x)

เปรียบเทียบ Side-by-Side Boxplot กับ Bar Chart ที่แสดงเฉพาะค่า Average

เราเห็นว่าไม่เพียงแต่ Average MEDV สำหรับ River-bounding Tracts สูงกว่า แต่ Distribution ทั้งหมดก็สูงกว่า

เรายังเห็นว่า River-bounding Tracts ทั้งหมดมี MEDV เหนือ $10K ซึ่งแตกต่างจาก Non-river-bounding Tracts

การใช้ Distribution Plots สำหรับ Attribute Derivation

Boxplots และ Histograms ที่ใช้กับ Numerical Attributes สามารถให้ทิศทางสำหรับการสร้าง Attributes ใหม่

ตัวอย่างเช่น พวกมันสามารถบอกวิธีการ Bin Numerical Attribute:

- Binning Numerical Target เพื่อใช้ Naive Bayes Classifier
- ในตัวอย่าง Boston Housing การเลือก Threshold เพื่อแปลง MEDV เป็น CAT.MEDV

Side-by-Side Boxplots

Boxplots มักจะถูกจัดเรียงเป็น Series โดยมี Plot ที่แตกต่างกันสำหรับแต่ละค่าของ Attribute ที่สอง ซึ่งแสดงบนแกน *x*

ตัวอย่างเช่น Side-by-Side Boxplot ใน Figure 3.3 เปรียบเทียบ MEDV สำหรับ Tracts ที่อยู่ใกล้แม่น้ำ (CHAS=1) กับที่ไม่ใกล้แม่น้ำ (CHAS=0)

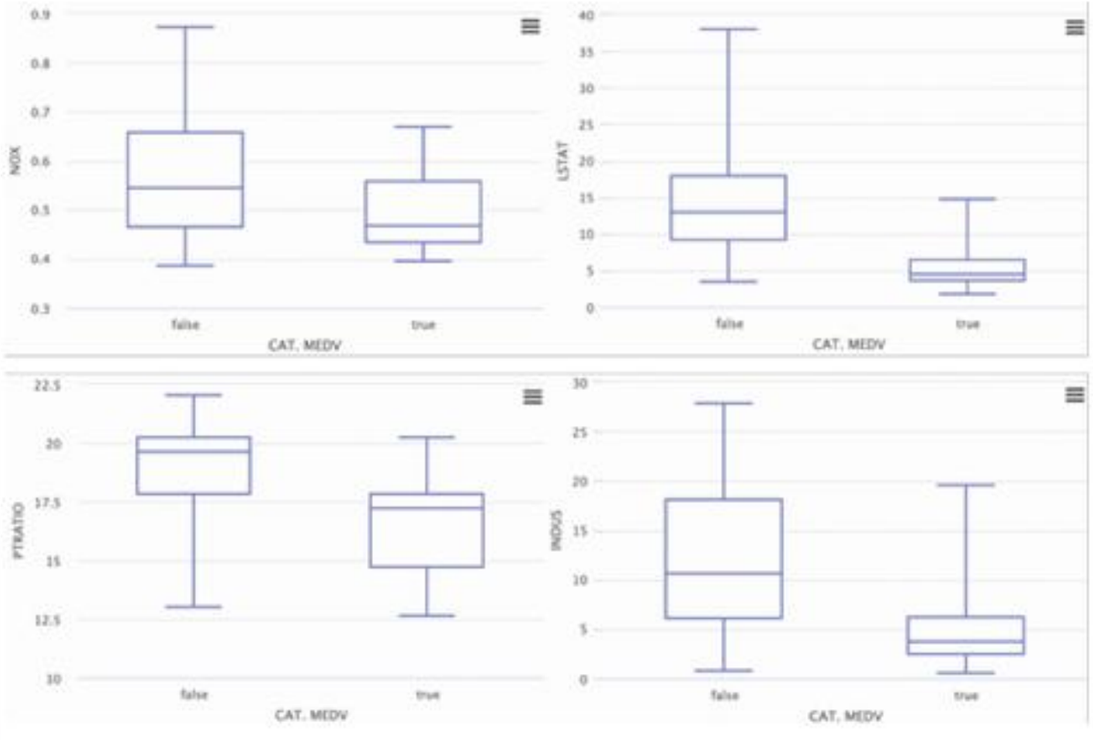
เราเห็นว่าไม่เพียงแต่ Average MEDV สำหรับ River-bounding Tracts สูงกว่า แต่ Distribution ทั้งหมดก็สูงกว่า (Median, Quartiles และ Min)

Side-by-Side Boxplots สำหรับ Classification

Side-by-Side Boxplots มีประโยชน์ใน Classification Tasks สำหรับการประเมินศักยภาพของ Numerical Predictors

ทำได้โดยใช้แกน *x* สำหรับ Categorical Target Attribute และแกน *y* สำหรับ Numerical Predictor

คู่ที่แยกออกจากกันมากที่สุด (เช่น PTRATIO และ INDUS) บอกถึง Predictors ที่มีประโยชน์ที่เป็นไปได้



Multidimensional Visualization

การขยาย Basic Charts ด้วย Features เช่น Color, Size และ Multiple Panels เพื่อแสดงข้อมูลหลายมิติ

Multidimensional Visualization

ใน Machine Learning ข้อมูลมักจะเป็น Multivariate โดยธรรมชาติ และ Analytics ถูกออกแบบมาเพื่อจับและวัด Multivariate Information

Visual Exploration ควรรวมแง่มุมสำคัญนี้ด้วย

ในส่วนถัดไป เราจะอธิบายวิธีการขยาย Basic Charts (และ Distribution Plots) ไปยัง Multidimensional Data Visualization โดยการเพิ่ม Features, ใช้ Manipulations และรวม Interactivity

Heatmaps: ภาพรวม

Heatmap เป็นการแสดงข้อมูล Numerical แบบกราฟิกที่ใช้สีเพื่อแสดงค่า

ในบริบทของ Machine Learning, Heatmaps มีประโยชน์เป็นพิเศษสำหรับสองวัตถุประสงค์:

Correlation Tables

Visualizing ตาราง Correlation

Missing Values

Visualizing Missing Values ในข้อมูล

โครงสร้างของ Heatmaps

ในทั้งสองกรณี ข้อมูลถูกถ่ายทอดในตารางสองมิติ:

Correlation Table

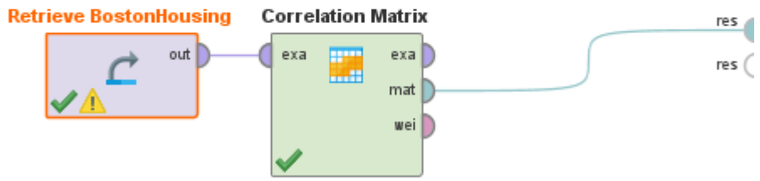
สำหรับ p Attributes มี p แถวและ p คอลัมน์

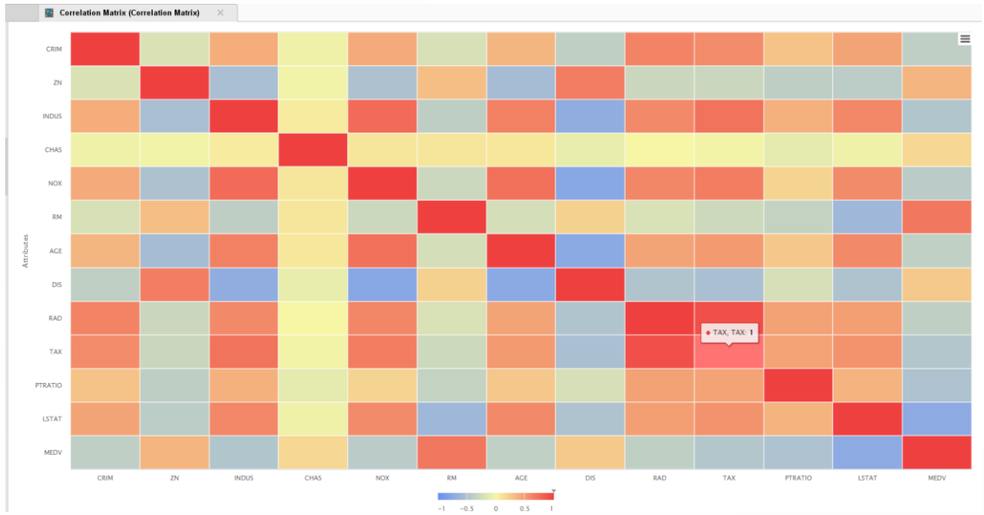
Data Table

ประกอบด้วย p คอลัมน์ (Attributes) และ n แถว (Observations)

หากจำนวนแถวมากมาย สามารถใช้ Subset ได้

ในทั้งสองกรณี การสแกน Color-coding นั้นง่ายและรวดเร็วกว่าการดูค่า



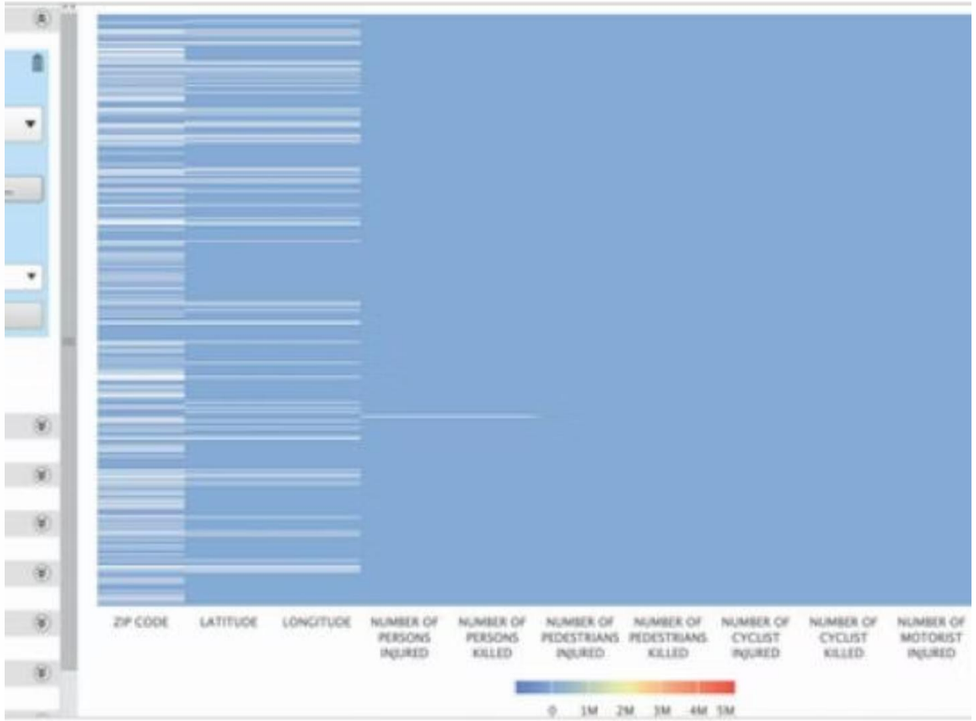


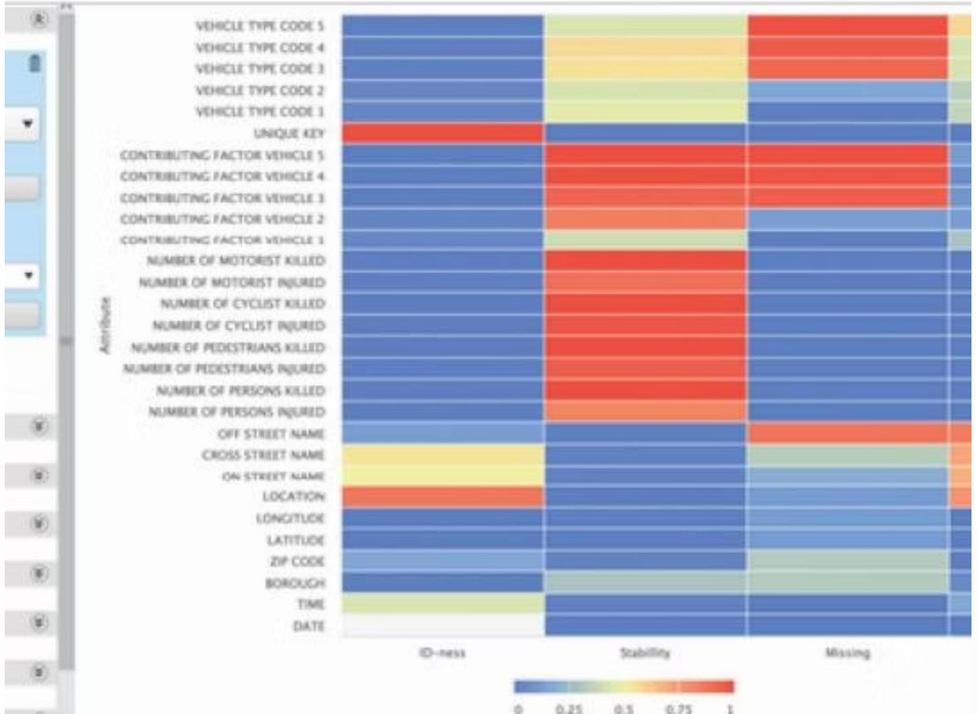
Correlation Table Heatmap

ตัวอย่างของ Correlation Table Heatmap แสดง Pairwise Correlations ทั้งหมดระหว่าง 13 Attributes (MEDV และ 12 Predictors)

เฉดสีที่เข้มขึ้นสอดคล้องกับ Correlation ที่แข็งแกร่งขึ้น (บวกหรือลบ) ทำให้ง่ายต่อการระบุ Correlations สูงและต่ำอย่างรวดเร็ว

การเพิ่มสี—ที่นี่สีแดงและสีน้ำเงิน—สามารถใช้เพื่อแยกแยะค่าบวกและลบ





Missing Value Heatmap

ใน Missing Value Heatmap แถวสอดคล้องกับ Records และคอลัมน์สอดคล้องกับ Attributes

ตัวอย่างแสดง Missing Value Heatmap สำหรับ Dataset เกี่ยวกับการชนกันของยานพาหนะ

Missing Value Cells ถูกแสดงด้วยสีขาวในขณะที่ส่วนที่เหลือแสดงเป็นสีน้ำเงิน

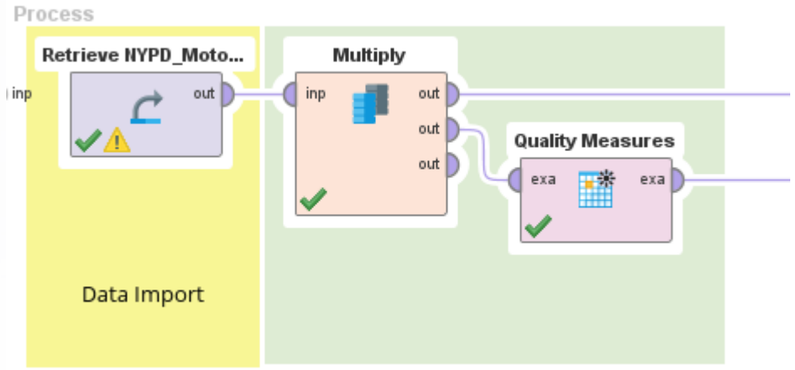
ประโยชน์ของ Missing Value Heatmap

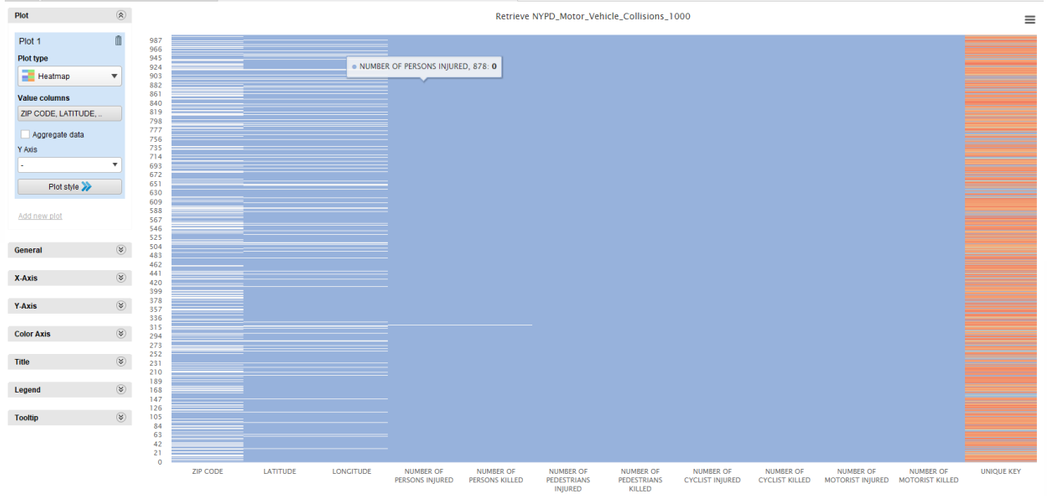
Missing Data Heatmap ช่วย Visualize ระดับและจำนวนของ "Missingness" ใน Dataset

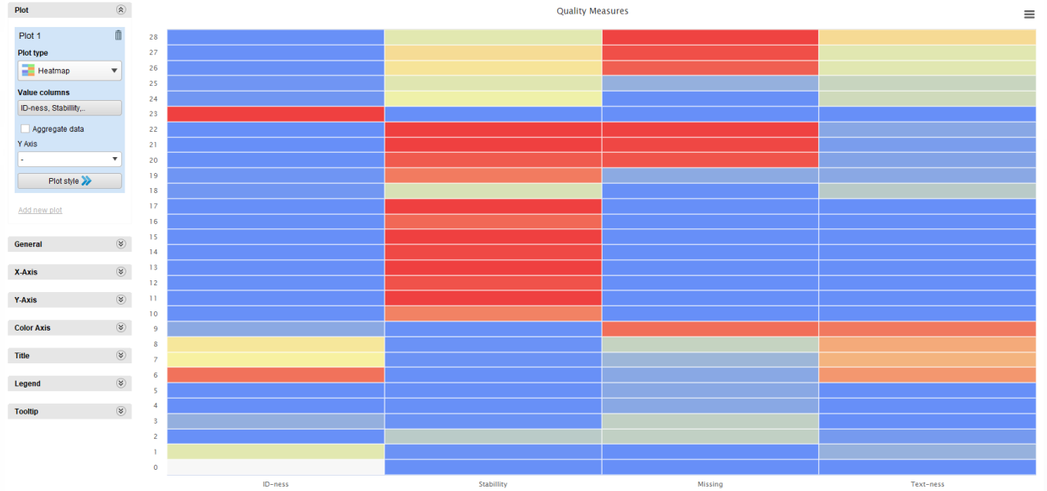
Patterns ของ "Missingness" บางอย่างปรากฏได้ง่าย:

- Attributes ที่หายไปเกือบทั้งหมดของ Observations
- Clusters ของแถวที่หายไปหลายค่า
- Attributes ที่มี Missingness น้อย

ข้อมูลนี้สามารถใช้สำหรับการกำหนดวิธีจัดการกับ Missingness (เช่น การลบบาง Attributes, การลบบาง Records, การ Impute หรือผ่านเทคนิคอื่นๆ)







Quality Measures Heatmap

Operator *Quality Measures* ของ RapidMiner คำนวณลักษณะคุณภาพที่มีประโยชน์สำหรับทุก Attribute ในข้อมูลบนมาตราส่วน 0–1:

- **ID-ness:** อัตราส่วนของจำนวนค่าที่แตกต่างกันต่อจำนวน Records ทั้งหมด
- **Stability:** อัตราส่วนของจำนวน Records ที่มีค่าที่พบบ่อยที่สุดต่อจำนวน Records ที่ไม่ Missing
- **Missing:** อัตราส่วนของจำนวน Missing Values ต่อจำนวน Records ทั้งหมด
- **Text-ness:** วัดระดับที่ Attribute มีข้อมูล Free Text

การเพิ่ม Features ใน Basic Charts

Basic Charts สามารถถ่ายทอดข้อมูลที่หลากหลายยิ่งขึ้นด้วย Features เช่น Color, Size และ Multiple Panels

และโดยการเปิดใช้งาน Operations เช่น Rescaling, Aggregation และ Interactivity

การเพิ่มเติมเหล่านี้ช่วยให้สามารถดูมากกว่าหนึ่งหรือสอง Attributes ในแต่ละครั้ง

Multidimensional Features

ความงามของการเพิ่มเติมเหล่านี้อยู่ที่ประสิทธิผลในการแสดงข้อมูลที่ซับซ้อนในลักษณะที่เข้าใจได้ง่าย

Features ที่มีประสิทธิภาพขึ้นอยู่กับความเข้าใจว่า Visual Perception ทำงานอย่างไร

วัตถุประสงค์คือการทำให้ข้อมูลเข้าใจได้มากขึ้น ไม่ใช่แค่การแสดงข้อมูลในมิติที่สูงขึ้น (เช่น Three-dimensional Plots ซึ่งมักจะเป็น Visualizations ที่ไม่มีประสิทธิภาพ)

การเพิ่ม Attributes: ประเภทของข้อมูล

เพื่อรวม Attributes เพิ่มเติมใน Plot เราต้องพิจารณาประเภทของ Attribute ที่จะรวม:

Categorical Information

วิธีที่ดีที่สุดคือการใช้ Hue, Shape หรือ Multiple Panels

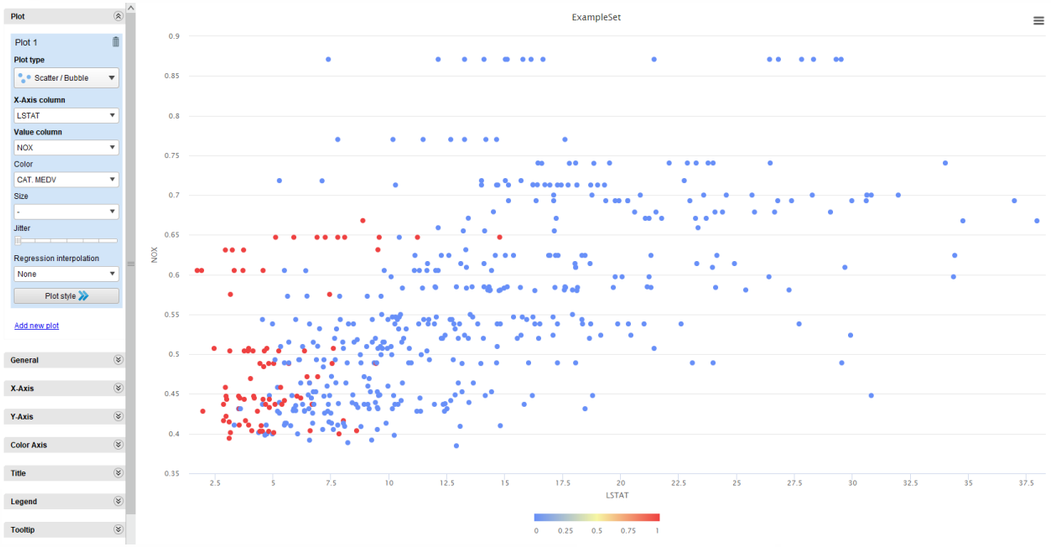
Numerical Information

เราสามารถใช้ Color Intensity หรือ Size

Temporal Information

สามารถเพิ่มผ่าน Animation

การใช้ Color-coded Scatter Plots สำหรับ Classification



การรวม Categorical และ/หรือ Numerical Attributes เพิ่มเติมเข้าไปใน Basic (และ Distribution) Charts หมายความว่าเราสามารถใช้ทั้งหมดสำหรับทั้ง Prediction และ Classification Tasks

Chart ที่มีประสิทธิภาพมากสำหรับ Classification คือ Scatter Plot ของ Numerical Predictors สองตัวที่ Color-coded โดย Categorical Target Attribute

Color-coding สำหรับ Prediction Tasks

ในบริบทของ Prediction, Color-coding สนับสนุนการสำรวจ Conditional Relationship ระหว่าง Numerical Target Attribute (บนแกน y) และ Numerical Predictor

Color-coded Scatter Plots ช่วยประเมินความจำเป็นในการสร้าง Interaction Terms

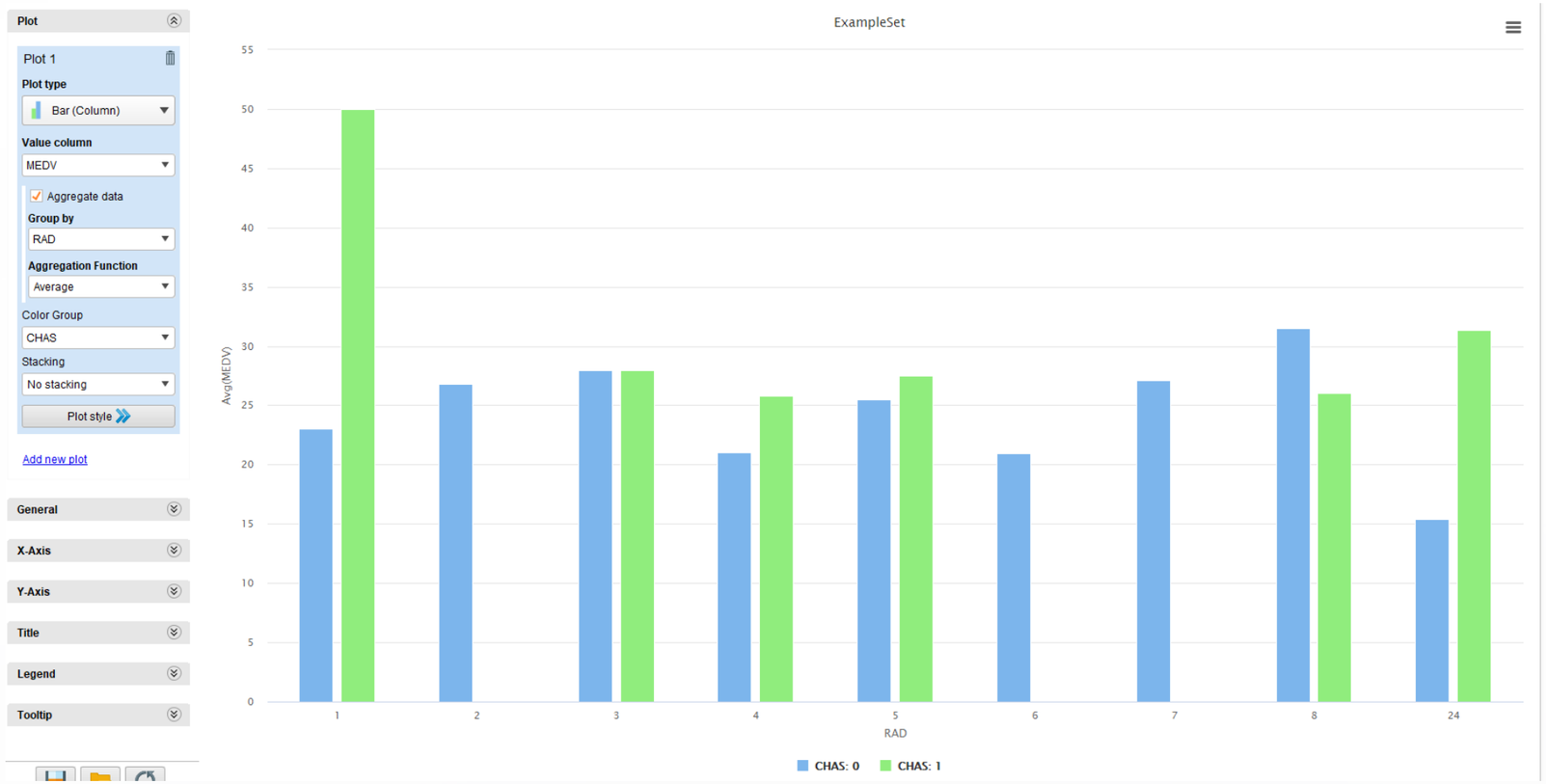
ตัวอย่างเช่น: Relationship ระหว่าง MEDV และ LSTAT แตกต่างกันหรือไม่สำหรับ Tracts ใกล้กับแม่น้ำเทียบกับที่ห่างจากแม่น้ำ?

การใช้ Color ใน Bar Charts

Color สามารถใช้เพื่อรวม Categorical Attributes เพิ่มเติมเข้าไปใน Bar Chart ตราบใดที่จำนวนหมวดหมู่มีน้อย

เมื่อจำนวนหมวดหมู่มีมาก ทางเลือกที่ดีกว่าคือการใช้ Multiple Panels

ตัวอย่างแสดง Bar Chart ของ Average MEDV โดย RAD ที่แบ่งออกเป็นหมวดหมู่ CHAS ด้วย Color Grouping



การสร้าง Multiple Panels (Trellising)

การสร้าง Multiple Panels (เรียกอีกอย่างว่า "Trellising") ทำได้โดยการแบ่ง Observations ตาม Categorical Attribute และสร้าง Chart แยกต่างหาก (ประเภทเดียวกัน) สำหรับแต่ละหมวดหมู่

หมายเหตุ: RapidMiner สนับสนุน Color Grouping โดย Attribute แต่ไม่สนับสนุนการสร้าง Multiple Panels

Insights จาก Color-coded Bar Charts

จาก Color-coded Bar Chart ของ MEDV โดย RAD และ CHAS เราเห็นว่า:

- Average MEDV สำหรับระดับ Highway Accessibility (RAD) ที่แตกต่างกันมีพฤติกรรมที่แตกต่างกันสำหรับ Tracts ใกล้แม่น้ำ (แถบสีเขียว) เมื่อเทียบกับ Tracts ห่างจากแม่น้ำ (แถบสีน้ำเงิน)
- สิ่งนี้เด่นชัดเป็นพิเศษสำหรับ RAD = 1
- ไม่มี Near-river Tracts ใน RAD Levels 2, 6 และ 7

ข้อมูลดังกล่าวอาจนำเราไปสู่การสร้าง Interaction Term ระหว่าง RAD และ CHAS

Scatter Plot Matrix

Chart พิเศษที่ใช้ Scatter Plots กับ Multiple Panels คือ **Scatter Plot Matrix**

ในนั้น Pairwise Scatter Plots ทั้งหมดถูกแสดงในการแสดงผลเดียว

Panels ใน Matrix Scatter Plot ถูกจัดระเบียบในลักษณะพิเศษ โดยแต่ละคอลัมน์และแต่ละแถวสอดคล้องกับ Attribute ดังนั้น Intersections จึงสร้าง Pairwise Scatter Plots ที่เป็นไปได้ทั้งหมด

การใช้ Scatter Plot Matrix

Scatter Plot Matrix มีประโยชน์ใน Unsupervised Learning สำหรับ:

- ศึกษา Associations ระหว่าง Numerical Attributes
- ตรวจจับ Outliers
- ระบุ Clusters

สำหรับ Supervised Learning สามารถใช้สำหรับ:

- ตรวจสอบ Pairwise Relationships (และลักษณะของพวกมัน) ระหว่าง Predictors
- สนับสนุน Attribute Transformations และ Attribute Selection
- สำหรับ Prediction แสดง Relationship ของ Outcome กับ Numerical Predictors

การเพิ่ม Attributes เพิ่มเติม

เมื่อ Hue ถูกใช้แล้ว Categorical Attributes เพิ่มเติมสามารถเพิ่มผ่าน Shape และ Multiple Panels

อย่างไรก็ตาม ต้องดำเนินการอย่างระมัดระวังในการเพิ่ม Multiple Attributes เพราะการแสดงผลอาจกลายเป็นที่แออัดเกินไป และ Visual Perception จะสูญหาย

การเพิ่ม Numerical Attribute ผ่าน Size มีประโยชน์โดยเฉพาะใน Scatter Plots (จึงสร้าง "Bubble Plots") เพราะใน Scatter Plot จุดแสดง Individual Observations

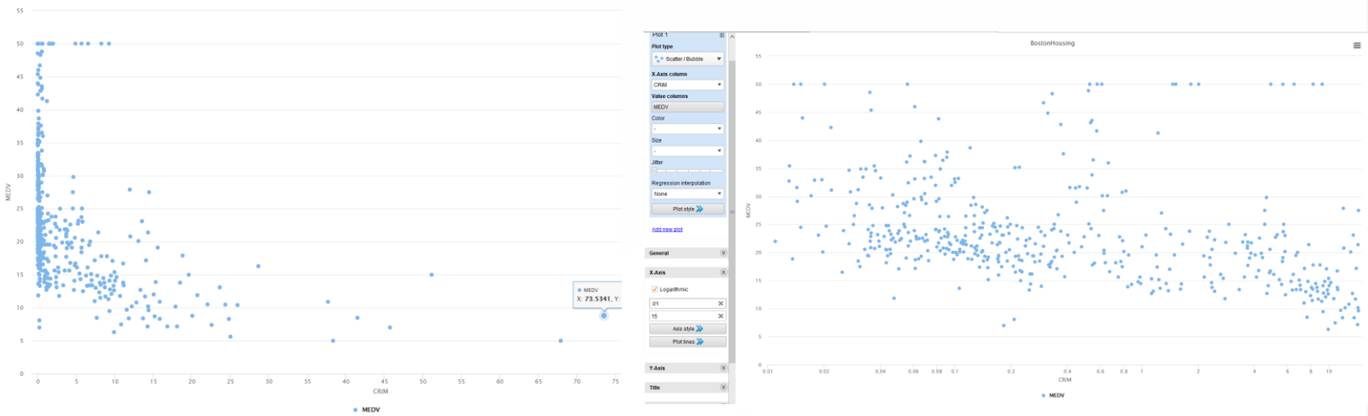
การเพิ่มมิติ Temporal: Animation

การเพิ่มมิติ Temporal ใน Chart เพื่อแสดงว่าข้อมูลเปลี่ยนแปลงอย่างไรตามเวลาสามารถทำได้ผ่าน Animation

ตัวอย่างที่มีชื่อเสียงคือ Animated Scatter Plots ของ Rosling ที่แสดงว่า World Demographics เปลี่ยนแปลงอย่างไรตลอดหลายปี (www.gapminder.org)

อย่างไรก็ตาม: ในขณะที่ Animations ประเภทนี้ใช้งานได้สำหรับ "Statistical Storytelling" แต่ไม่มีประสิทธิภาพมากสำหรับ Data Exploration

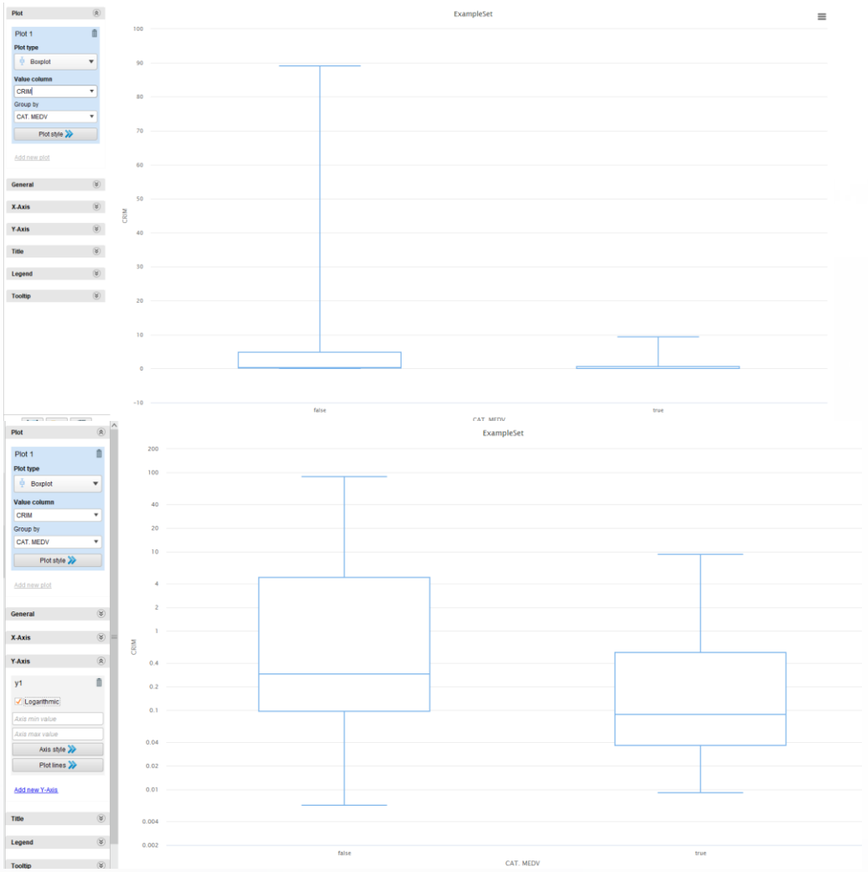
Manipulations: Rescaling



การเปลี่ยน Scale ในการแสดงผลสามารถปรับปรุง Chart และทำให้ Relationships ชัดเจนขึ้น

ตัวอย่างเช่น ในรูป เราเห็นผลของการเปลี่ยนทั้งสองแกนของ Scatter Plot (บน) และแกน *y* ของ Boxplot (ล่าง) เป็น Logarithmic (Log) Scale

ใน RapidMiner ทำได้โดยการเลือกตัวเลือก *Logarithmic* สำหรับแต่ละแกนใน Plot Settings



ประโยชน์ของ Rescaling

ในขณะที่ Charts ต้นฉบับ (ซ้าย) ยากต่อการเข้าใจ Patterns กลับมองเห็นได้ใน Log Scale (ขวา)

Scatter Plots

ลักษณะของ Relationship ระหว่าง MEDV และ CRIM ยากต่อการกำหนดใน Original Scale เพราะจุดมากเกินไป "แออัด" ใกล้แกน y

Rescaling ขจัดความแออัดนี้และช่วยให้มองเห็น Linear Relationship ระหว่าง Log-scaled Attributes ทั้งสองได้ดีขึ้น

Boxplot

การแสดงความแออัดไปทางแกน x ใน Original Units ไม่ช่วยให้เราเปรียบเทียบขนาด Box ทั้งสอง ตำแหน่ง Lower Outliers และข้อมูล Distribution ส่วนใหญ่

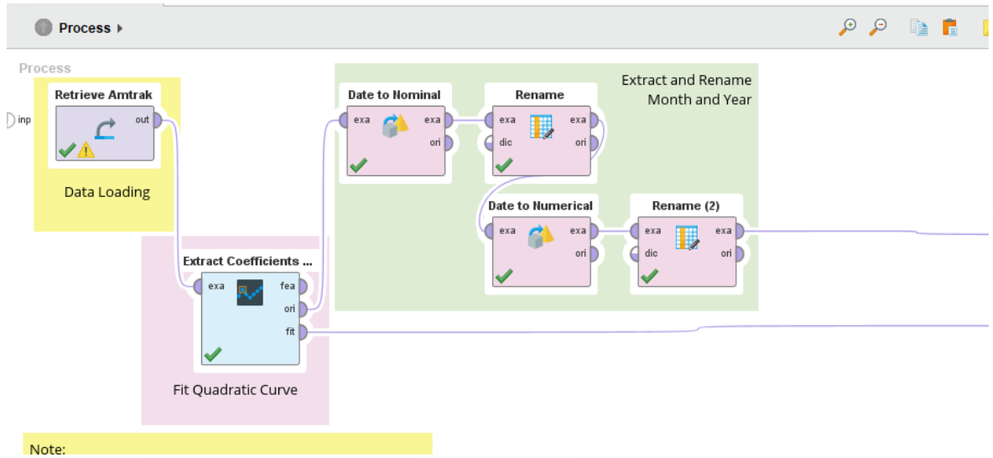
Rescaling ขจัดผล "Crowding to the x-axis" จึงช่วยให้เปรียบเทียบ Boxplots ทั้งสองได้

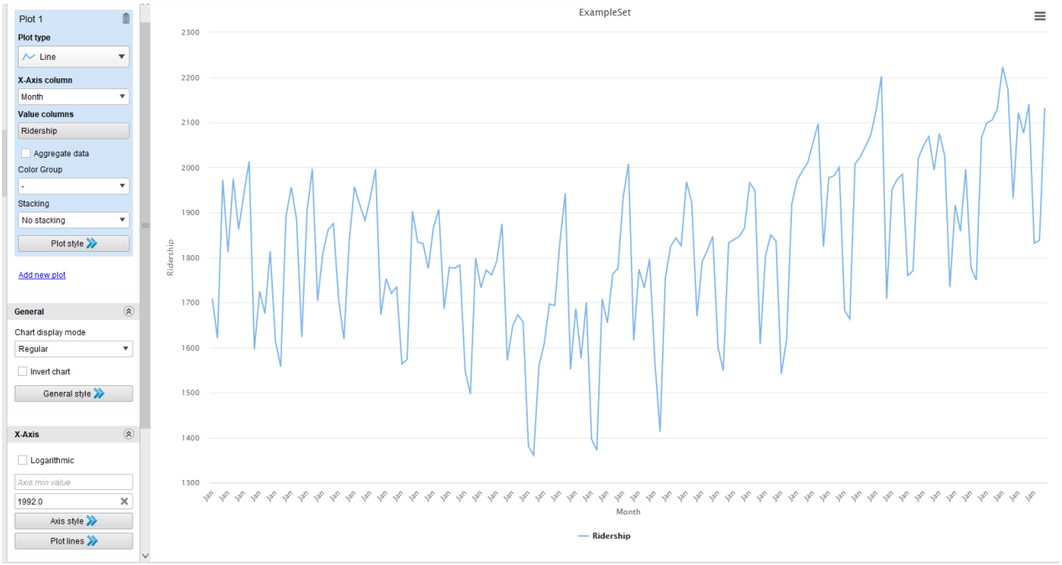
Aggregation และ Hierarchies

Manipulation ที่มีประโยชน์อีกอย่างของ Scaling คือการเปลี่ยนระดับของ Aggregation

สำหรับ Temporal Scale เราสามารถ Aggregate ตาม Granularity ที่แตกต่างกัน (เช่น รายเดือน รายวัน รายชั่วโมง) หรือแม้แต่ตาม "Seasonal" Factor ที่สนใจเช่น Month of Year หรือ Day of Week

Aggregation ที่นิยมสำหรับ Time Series คือ Moving Average ซึ่ง Average ของค่าที่อยู่ใกล้เคียงภายใน Window Size ที่กำหนดถูก Plot





Zooming, Panning และ Filtering

Zooming และ Panning

ความสามารถในการ Zoom เข้าและออกจากพื้นที่บางส่วนของข้อมูลบน Plot มีความสำคัญสำหรับการเปิดเผย Patterns และ Outliers

Panning หมายถึง Operation ของการเคลื่อนย้าย Zoom Window ไปยังพื้นที่อื่นๆ

Filtering

Filtering หมายถึงการลบบาง Observations ออกจาก Visualization

วัตถุประสงค์ของ Filtering คือการมุ่งความสนใจไปที่ข้อมูลบางส่วนในขณะที่กำจัด "Noise" ที่สร้างโดยข้อมูลอื่นๆ

Zooming, Panning และ Filtering สนับสนุน Supervised และ Unsupervised Learning โดยช่วยระบุ Local Behavior ที่แตกต่างหรือผิดปกติ

สรุป: Visualizations และ Operations หลัก

บทนี้นำเสนอชุดของ Plots ที่สามารถใช้เพื่อสำรวจลักษณะหลายมิติของ Dataset

3

Basic Charts

Bar Charts, Line Charts และ Scatter Plots

2

Distribution Plots

Boxplots และ Histograms

5+

Enhancements

Color, Size, Multiple Panels, Rescaling, Aggregation, Zooming, Filtering

เรามุ่งเน้นว่า Visualizations และ Operations ที่แตกต่างกันสามารถสนับสนุน Machine Learning Tasks อย่างไร ตั้งแต่ Supervised Tasks (Prediction, Classification และ Time Series Forecasting) ไปจนถึง Unsupervised Tasks

Classwork

3.1 Shipments of Household Appliances: Line Charts. The file *ApplianceShipments.csv* contains the series of quarterly shipments (in millions of dollars) of US household appliances between 1985 and 1989.

a. Create a well-formatted time plot of the data using RapidMiner.

b. Does there appear to be a quarterly pattern? For a closer view of the patterns, zoom in to the range of 3500–5000 on the y-axis.

c. Using RapidMiner, create one chart with four separate lines, one line for each of Q1, Q2, Q3, and Q4. In RapidMiner, this can be achieved by adding attributes for quarter and year using the *Generate Attributes* operator. [*Hint*: The function expressions *prefix(Quarter,2)* and *suffix(Quarter,4)* can be used to create separate attributes for quarter and year, respectively.] Then, in your visualization, using Shipments as the value column, create a line chart aggregating the data with the *average* aggregation function, grouping it by year and then color-grouping it by quarter. This will plot shipment vs. year for each quarter as a separate series on a line chart. Zoom in to the range of 3500–5000 on the y-axis. Does there appear to be a difference between quarters?

d. Using RapidMiner, create a line chart of the series at a yearly aggregated level (i.e., the total shipments in each year).

Classwork

3.2 Sales of Riding Mowers: Scatter Plots. A company that manufactures riding mowers wants to identify the best sales prospects for an intensive sales campaign. In particular, the manufacturer is interested in classifying households as prospective owners or nonowners on the basis of Income (in \$1000s) and Lot Size (in 1000 ft^2). The marketing expert looked at a random sample of 24 households, given in the file *Riding-Mowers.csv*.

Using RapidMiner, create a scatter plot of Lot Size vs. Income, color-coded by the target attribute owner/nonowner.

Classwork

3.3 Laptop Sales at a London Computer Chain: Bar Charts and Boxplots. The file *LaptopSalesJanuary2008.csv* contains data for all sales of laptops at a computer chain in London in January 2008. This is a subset of the full dataset that includes data for the entire year.

a. Create a bar chart, showing the average retail price by store. Which store has the highest average? Which has the lowest?

b. To better compare retail prices across stores, create side-by-side boxplots of retail price by store. Now compare the prices in the two stores from (a). Does there seem to be a difference between their price distributions?