Chapter 3 – Data Visualization 数据可视化

Instructor: Zach Zhizhong ZHOU, Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

Graphs for Data Exploration 用于探索数据的图表

Basic Plots 基本图

Line Graphs 线图

Bar Charts 柱状图

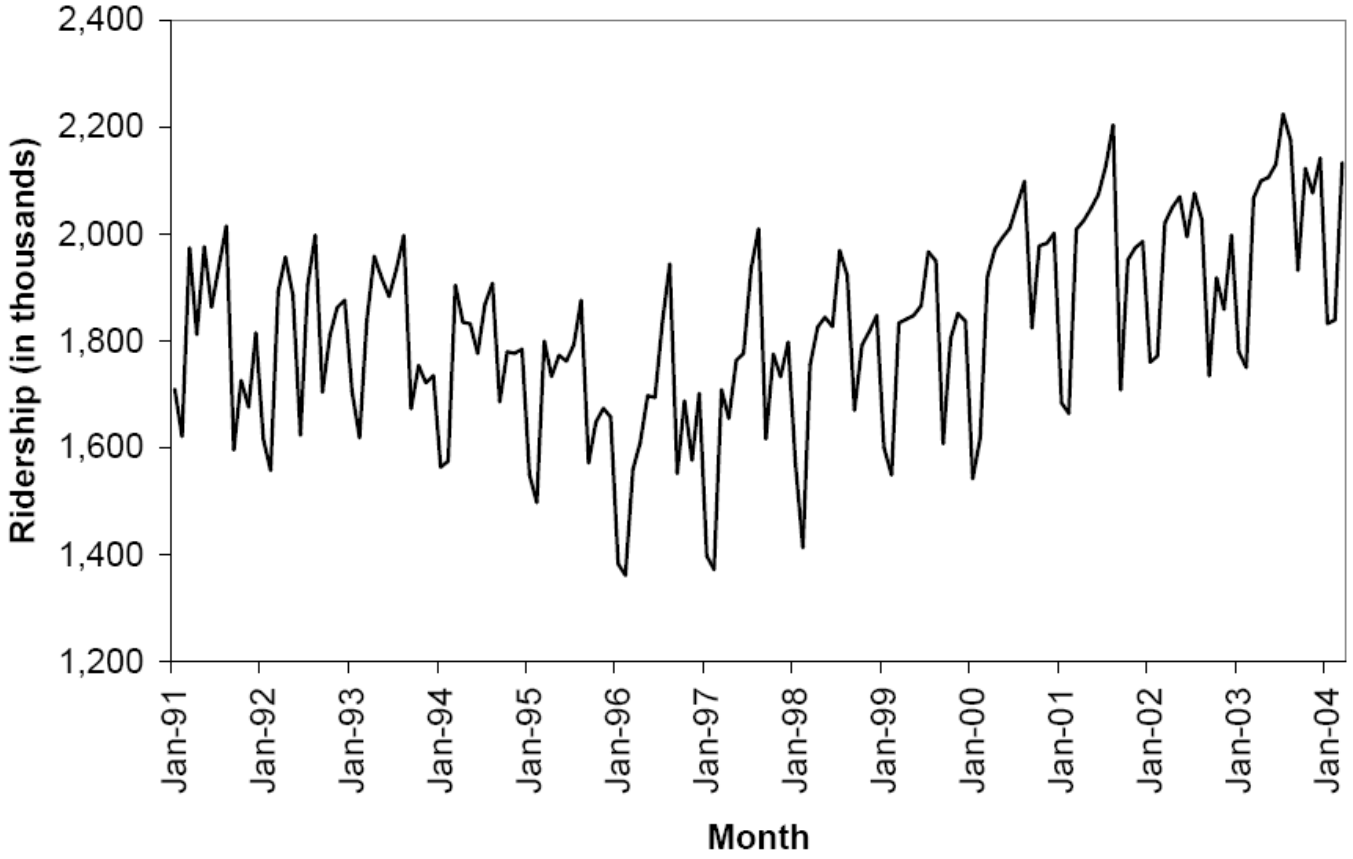
Scatterplots 散点图

Distribution Plots 分布图

Boxplots 盒状图

Histograms 直方图

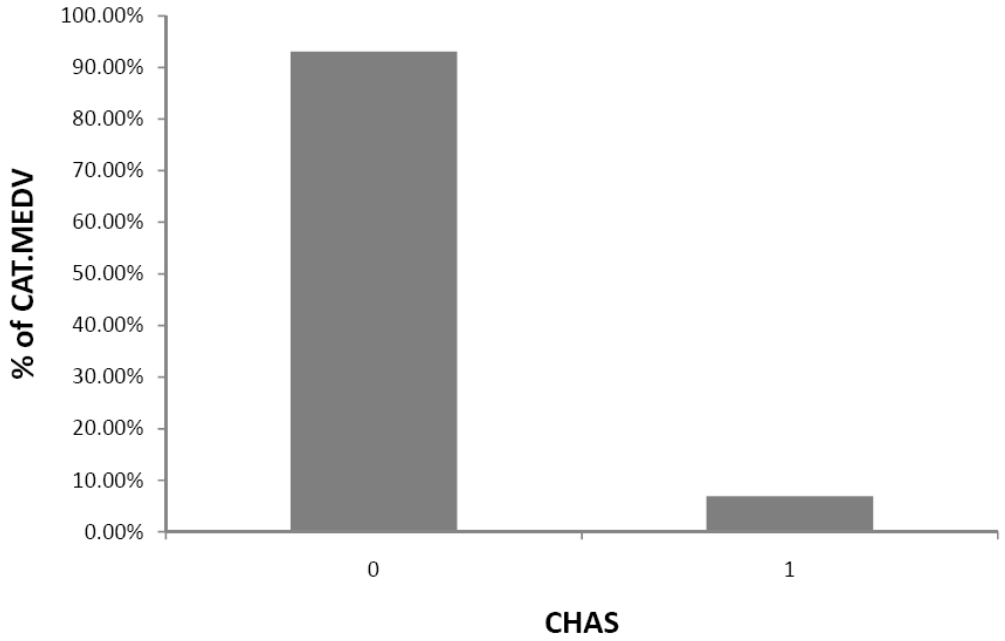
Line Graph for Time Series 时间序列数据的曲线图



Bar Chart for Categorical Variable 分类型变量的柱状图

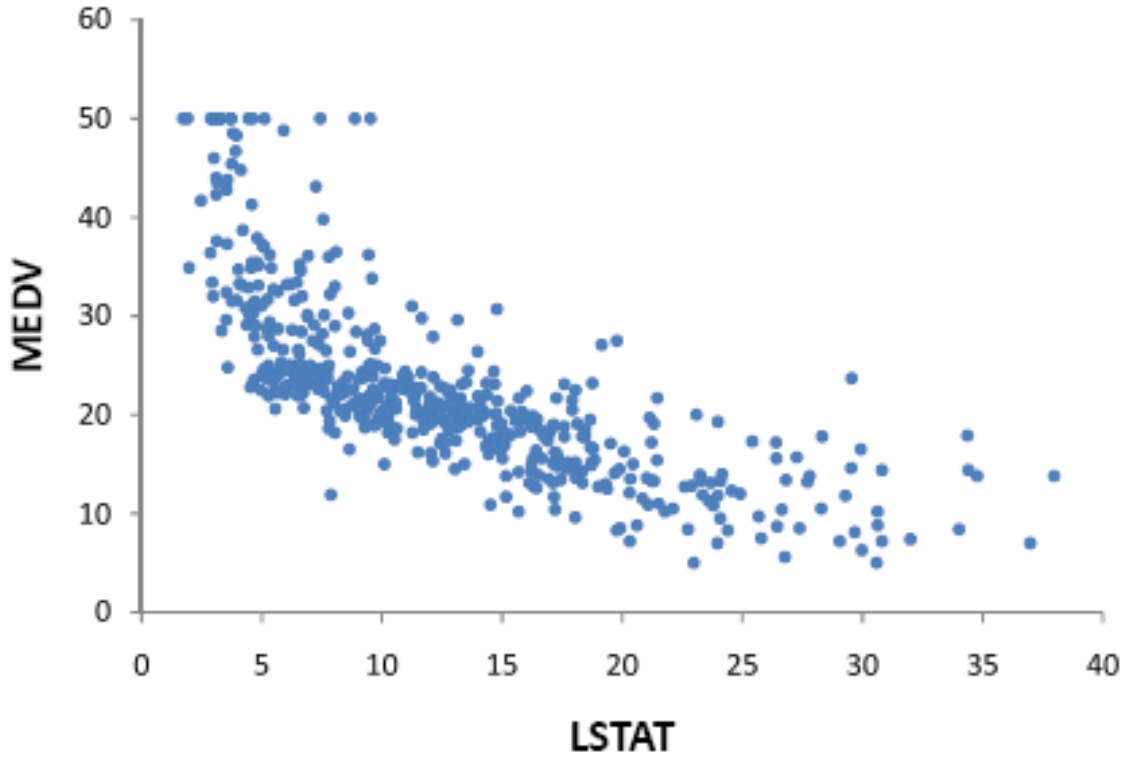
95% of tracts do not border Charles River

95%的地面与查尔斯河不相邻



Scatterplot 散点图

Displays relationship between two numerical variables 展示两个数值型变量之间的关系



Distribution Plots 分布图

- Display “how many” of each value occur in a data set

  展示取值在一个数据集中出现的次数

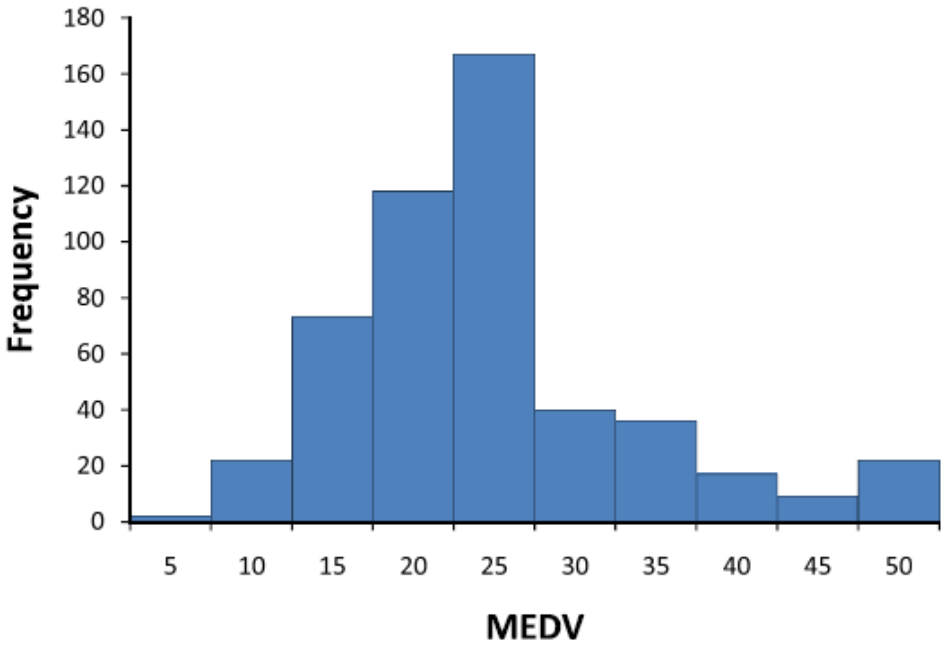
- Or, for continuous data or data with many possible values, “how many” values are in each of a series of ranges or “bins”

或者，对于连续型数据或者有很多可能取值的数据，出现在某个区间的数值的次数

Histograms 直方图

Histogram shows the distribution of the outcome variable (median house value) 直方图展示了结果变量的分布情况

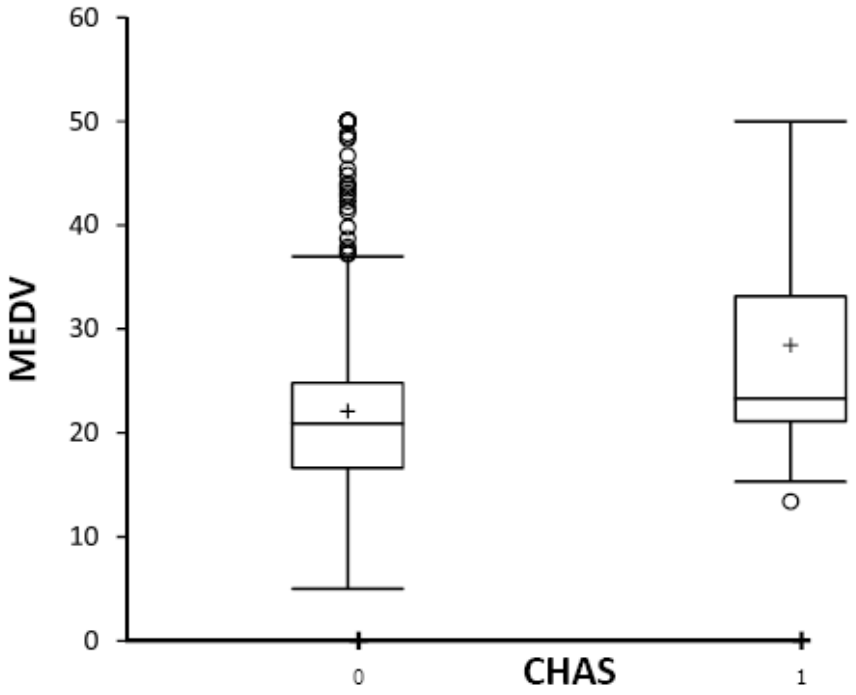
Boston Housing example:



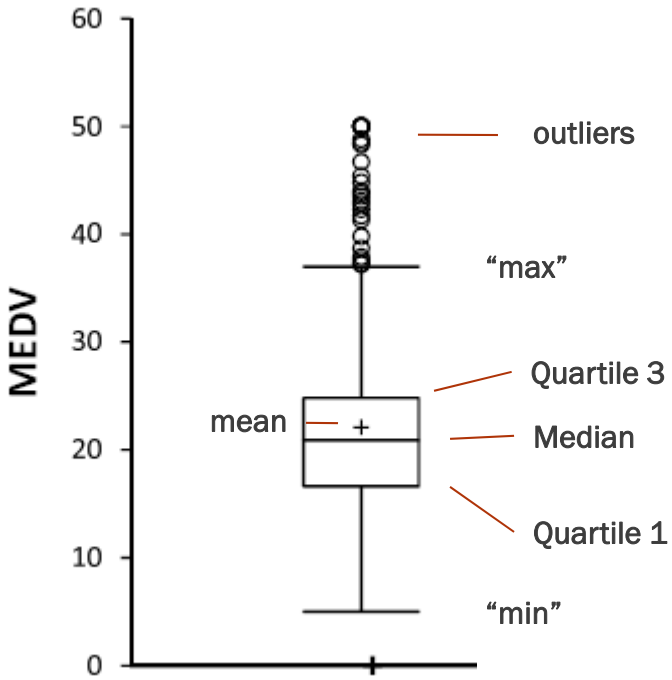
Boxplots 盒状图

Boston Housing Example: Display distribution of outcome variable (MEDV) for neighborhoods on Charles river (1) and not on Charles river (0) 展示结果变量（房屋价值中位数）在沿着查尔斯河以及不沿着查尔斯河2组数据中的分布情况

Side-by-side boxplots are useful for comparing subgroups 相邻的盒状图可用于比较2个数据子集



Box Plot 盒状图



- Top outliers defined as those above Q3+1.5(Q3-Q1).
- “max” = maximum of non-outliers
- Analogous definitions for bottom outliers and for “min”
- Details may differ across software

Heat Maps 热度图

Color conveys information 使用颜色传递信息

In data mining, used to visualize 在数据挖掘中，用于可视化以下信息

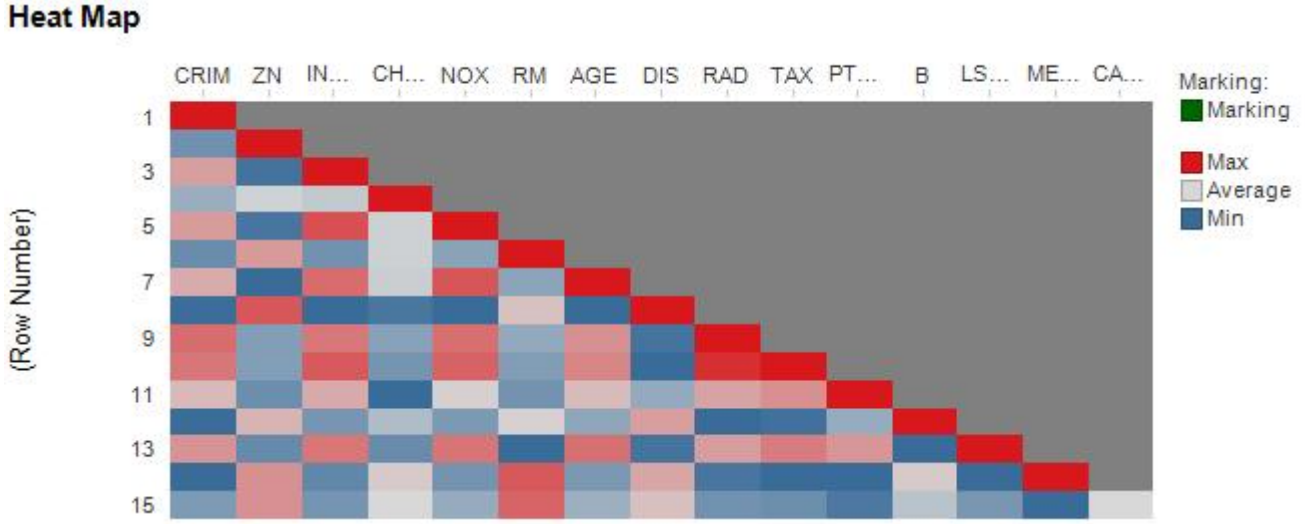
Correlations 相关关系

Missing Data 缺失数据

Heatmap to highlight correlations (Boston Housing) 热度图用于突出相关关系

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM														
ZN	-0.20													
INDUS	0.41	-0.53												
CHAS	-0.06	-0.04	0.06											
NOX	0.42	-0.52	0.76	0.09										
RM	-0.22	0.31	-0.39	0.09	-0.30									
AGE	0.35	-0.57	0.64	0.09	0.73	-0.24								
DIS	-0.38	0.66	-0.71	-0.10	-0.77	0.21	-0.75							
RAD	0.63	-0.31	0.60	-0.01	0.61	-0.21	0.46	-0.49						
TAX	0.58	-0.31	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91					
PTRATIO	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46				
B	-0.39	0.18	-0.36	0.05	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18			
LSTAT	0.46	-0.41	0.60	-0.05	0.59	-0.61	0.60	-0.50	0.49	0.54	0.37	-0.37		
MEDV	-0.39	0.36	-0.48	0.18	-0.43	0.70	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74	

In Excel (using conditional formatting)



In Spotfire

Multidimensional Visualization
多维可视化

Scatterplot with color added 上色的散点图

Boston Housing

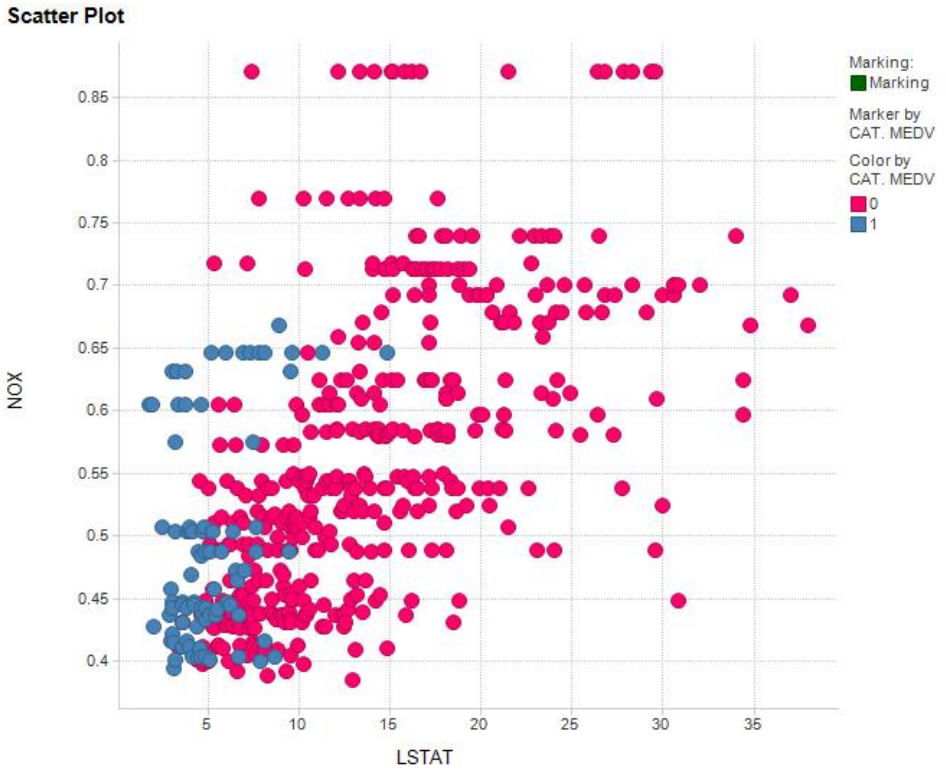
NOX vs. LSTAT

Red = low median value

Blue = high median value

NOX: nitric oxides concentration

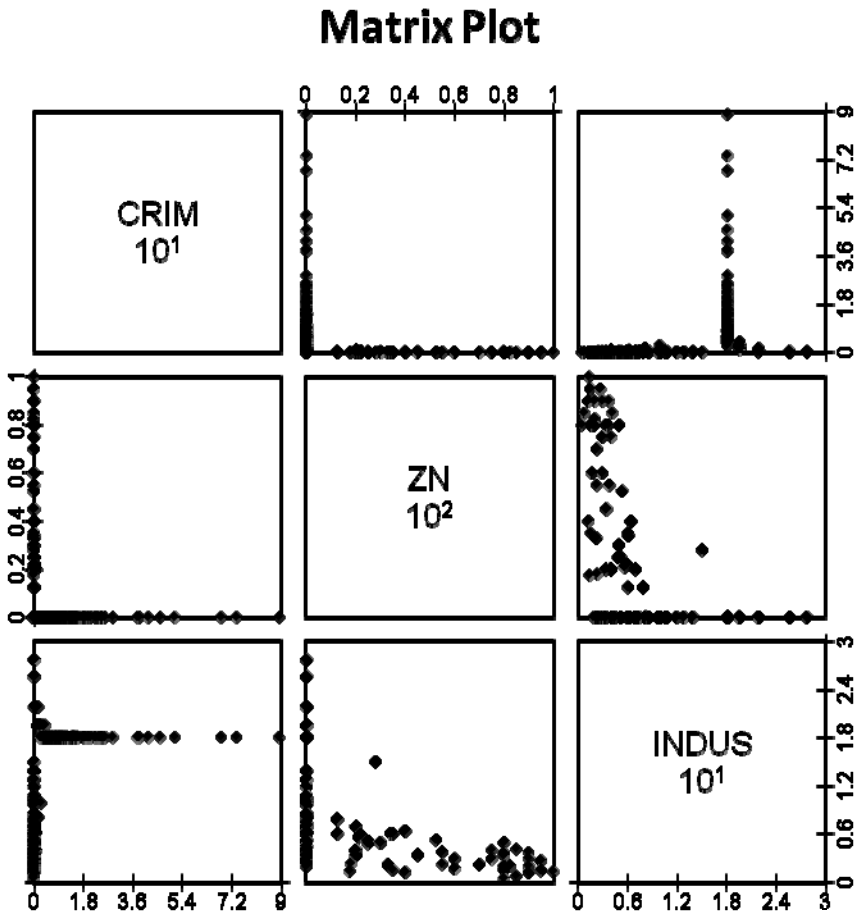
LSTAT: % low status population



Matrix Plot 矩阵图

Shows scatterplots for variable pairs 展示变量两两之间关系的散点图

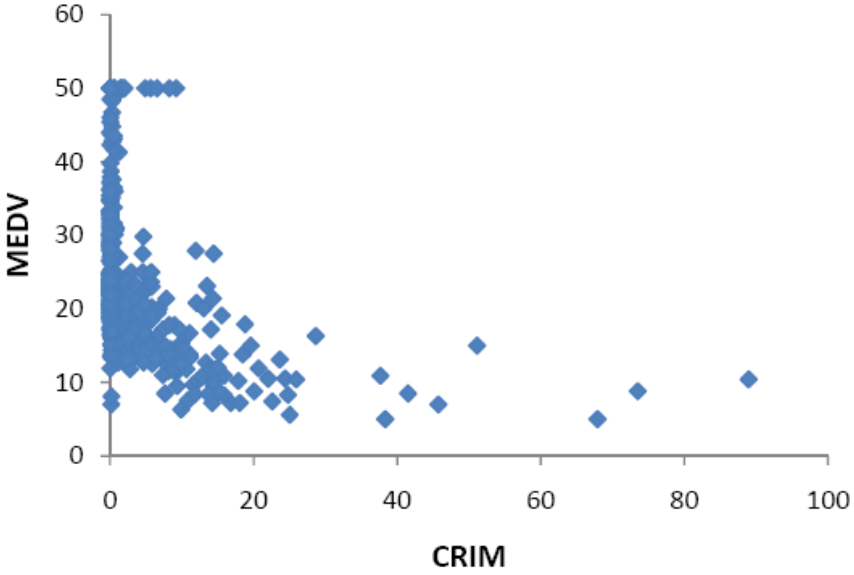
Example: scatterplots for 3 Boston Housing variables

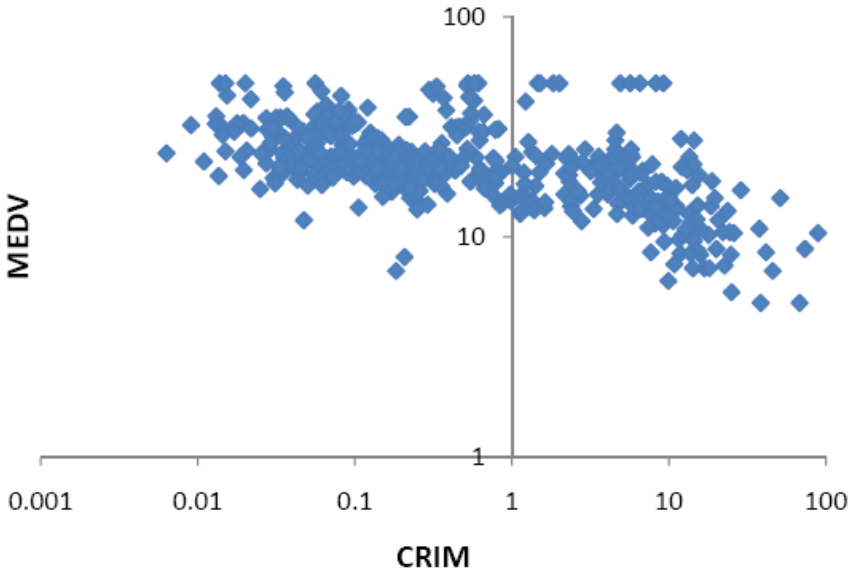


CRIM: per capita crime rate by town
ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS: proportion of non-retail business acres per town.

Rescaling to log scale (on right)

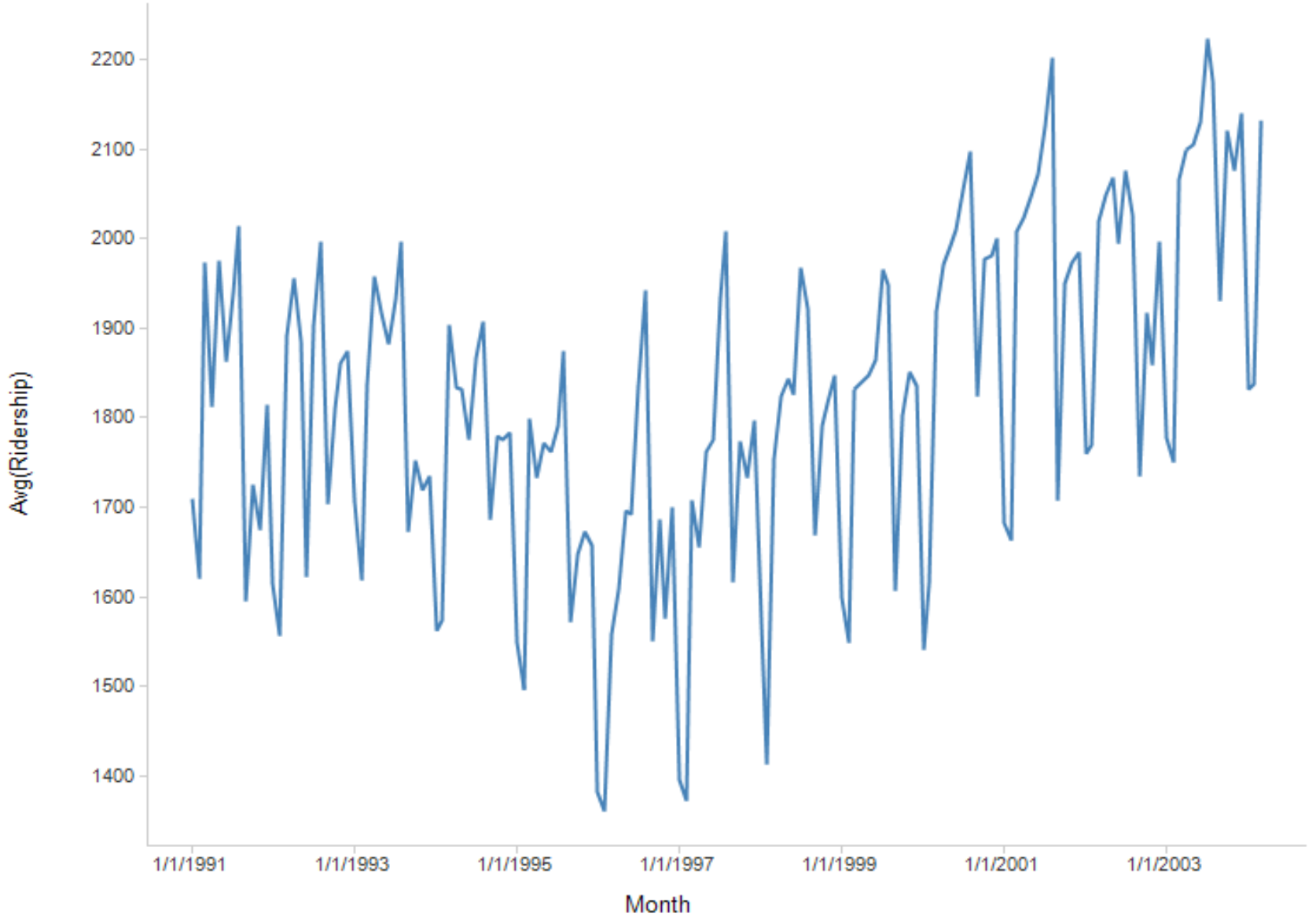
“uncrowds” the data 重新标度为log（取值）打散数据



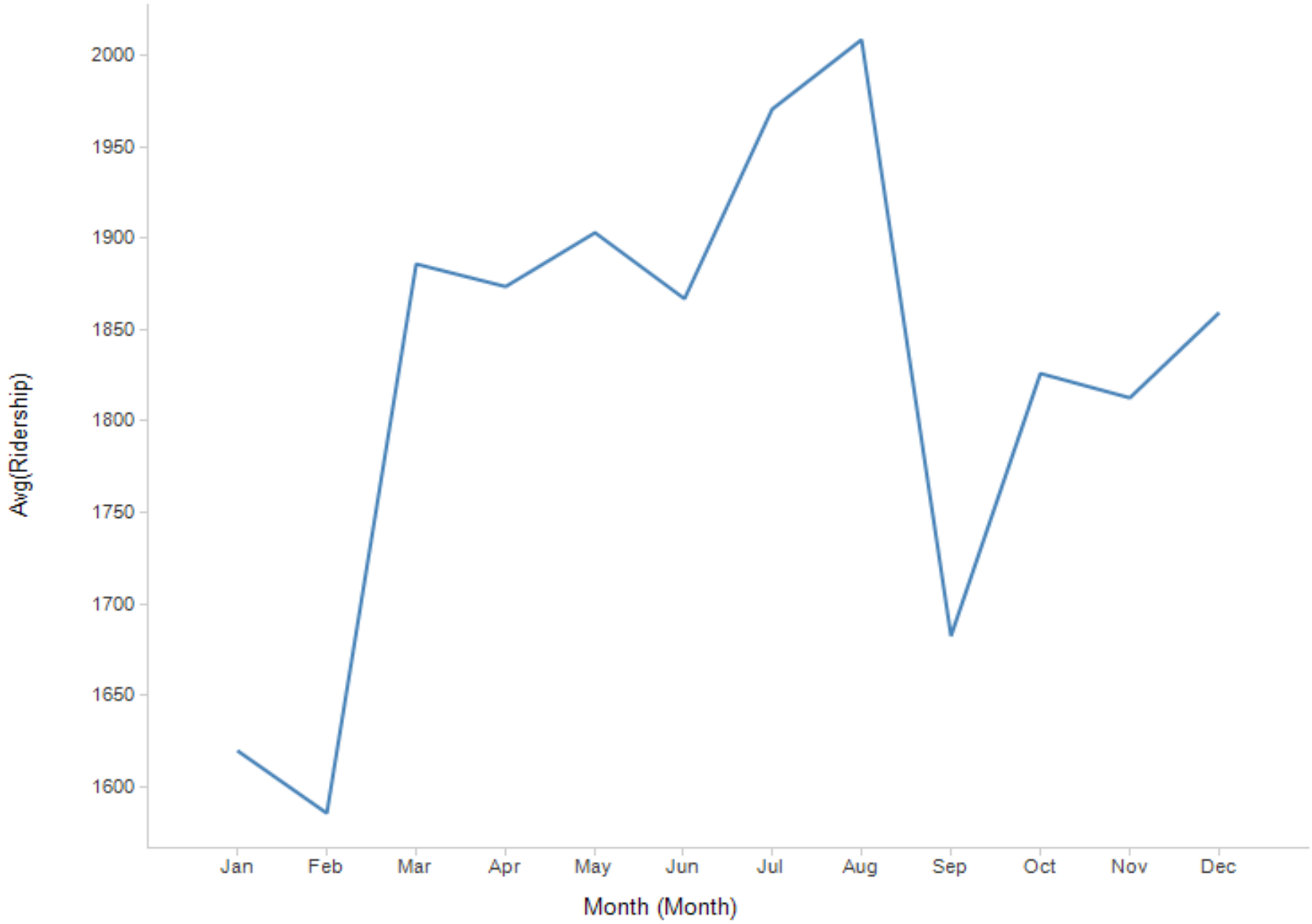


Aggregation 聚集

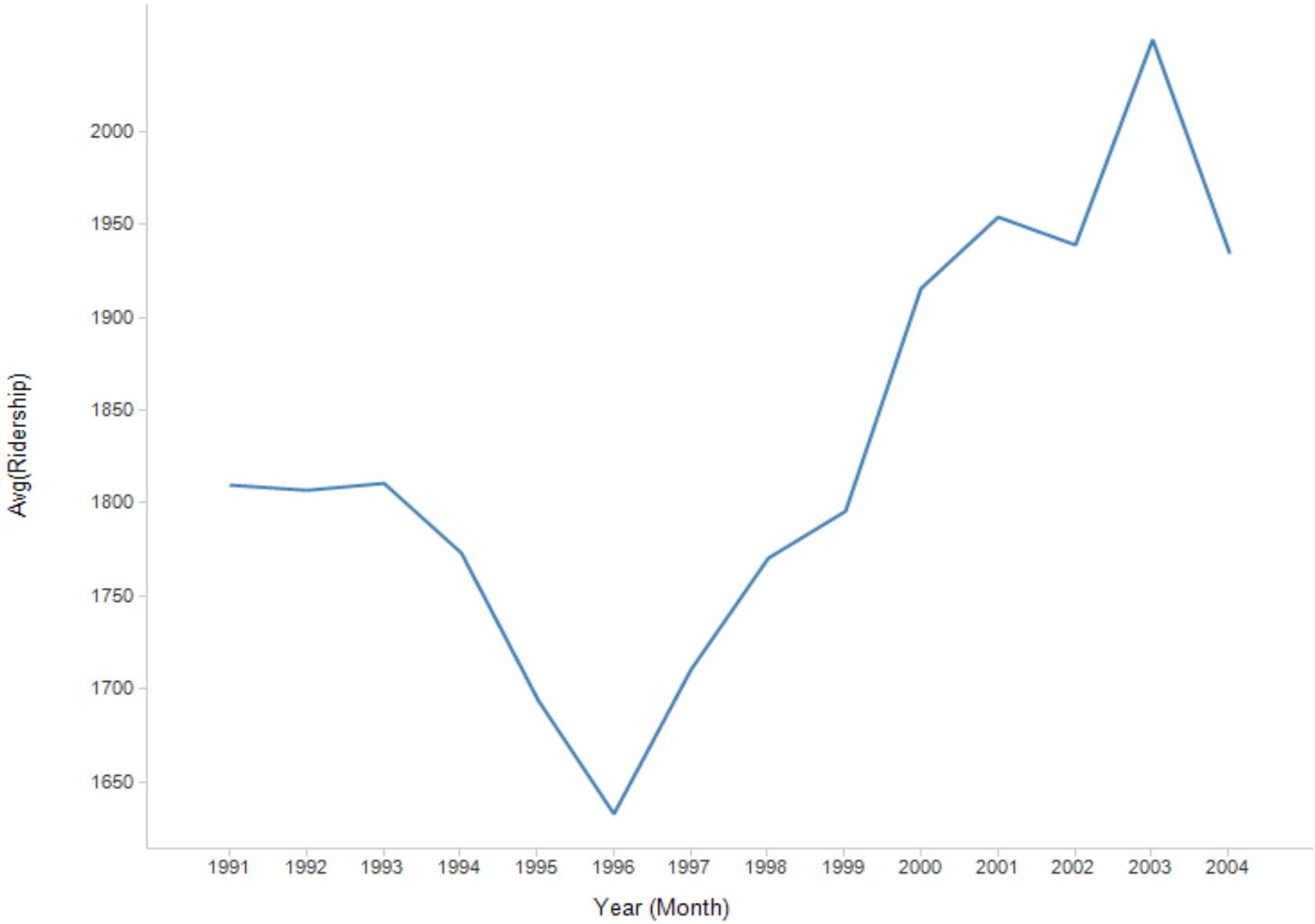
Amtrak Ridership – Monthly Data Amtrak 乘车人数 月数据



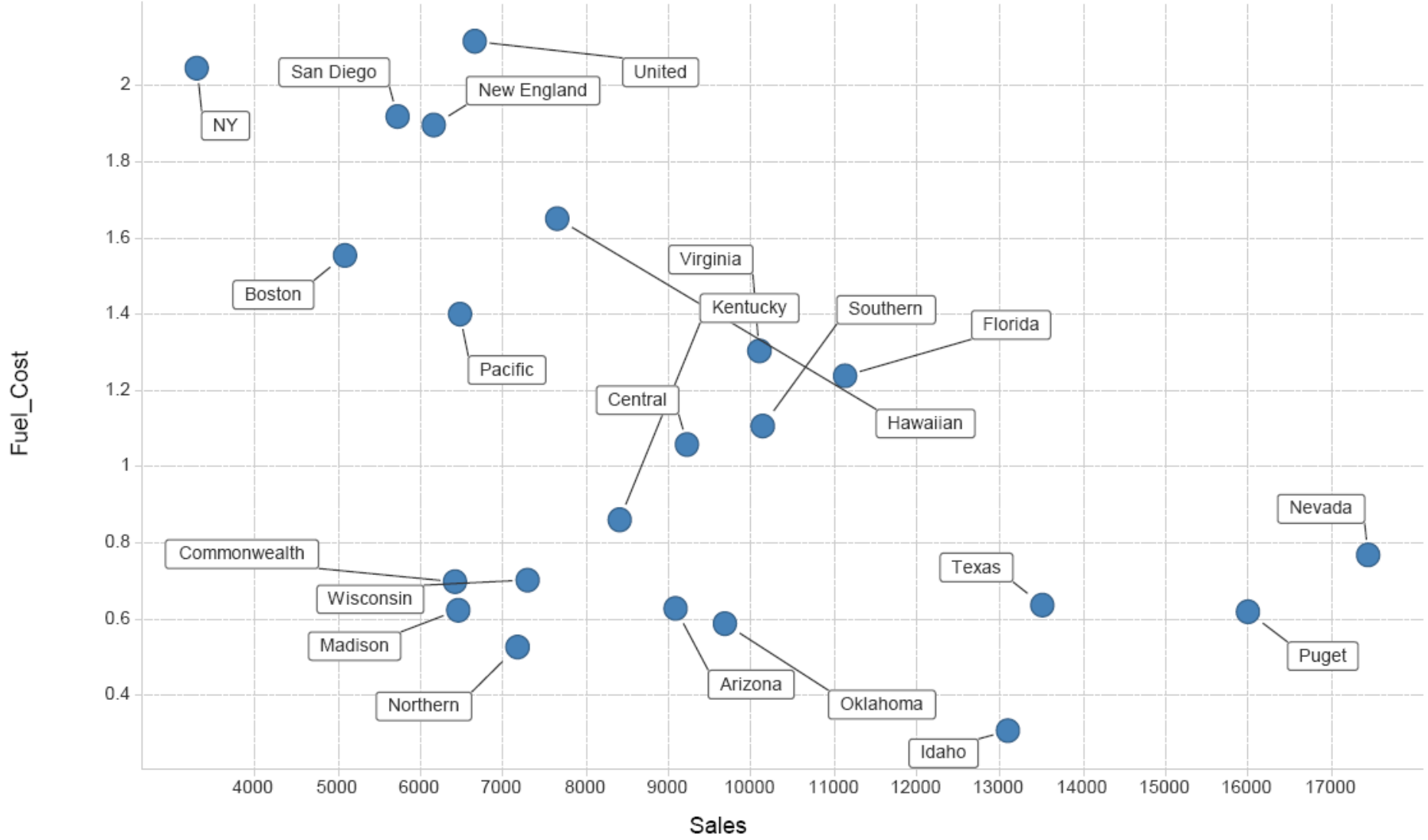
Aggregation – Monthly Average 聚集 - 按月平均



Aggregation – Yearly Average 聚集 - 按年平均



Scatter Plot with Labels (Utilities) 带标签的散点图 (水电煤气)



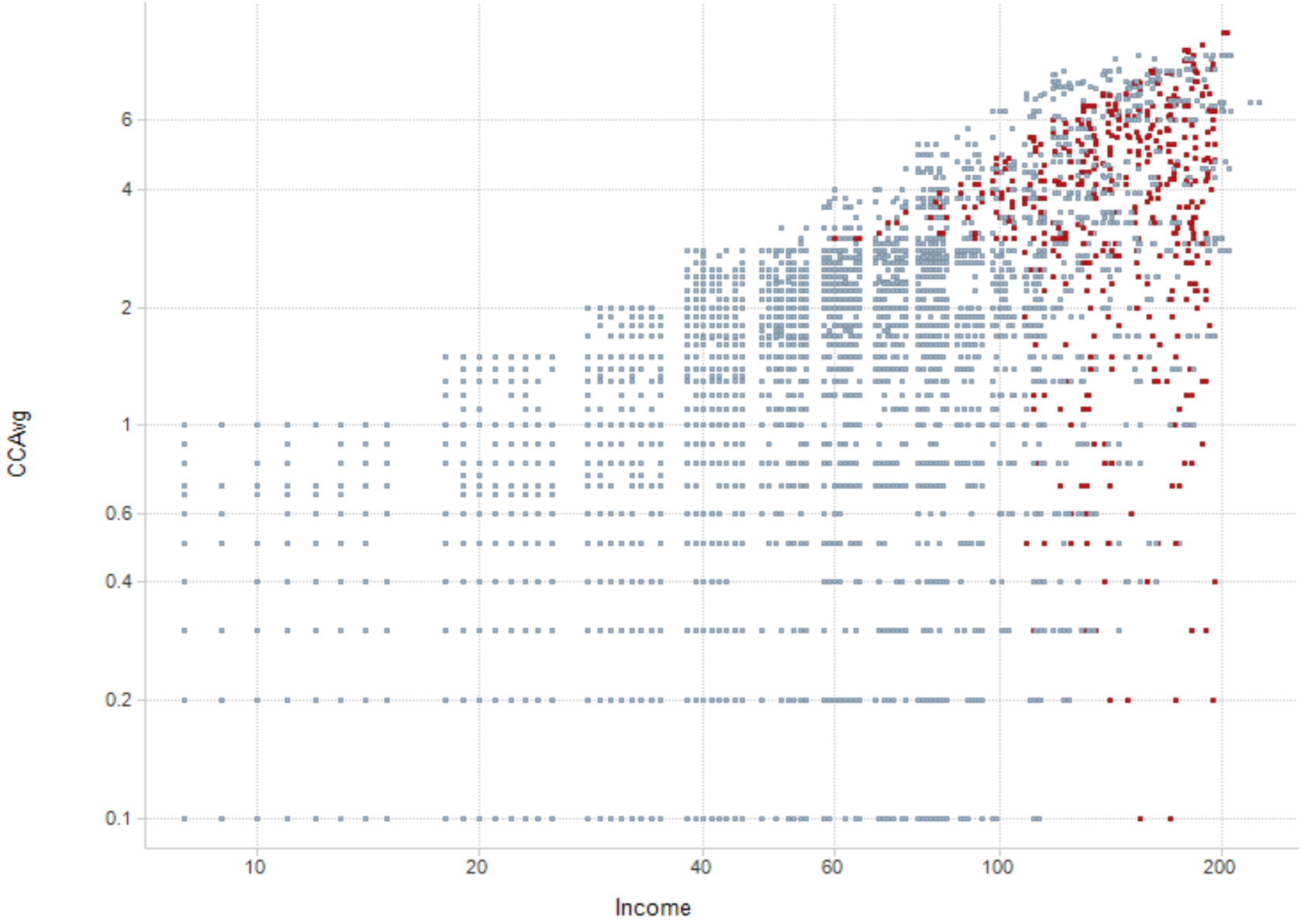
Scaling: Smaller markers, jittering, color contrast (Universal Bank; red = accept loan)

Jittering 抖动

- Moving markers by a small random amount 将图上的记号随机稍稍移动一段距离。

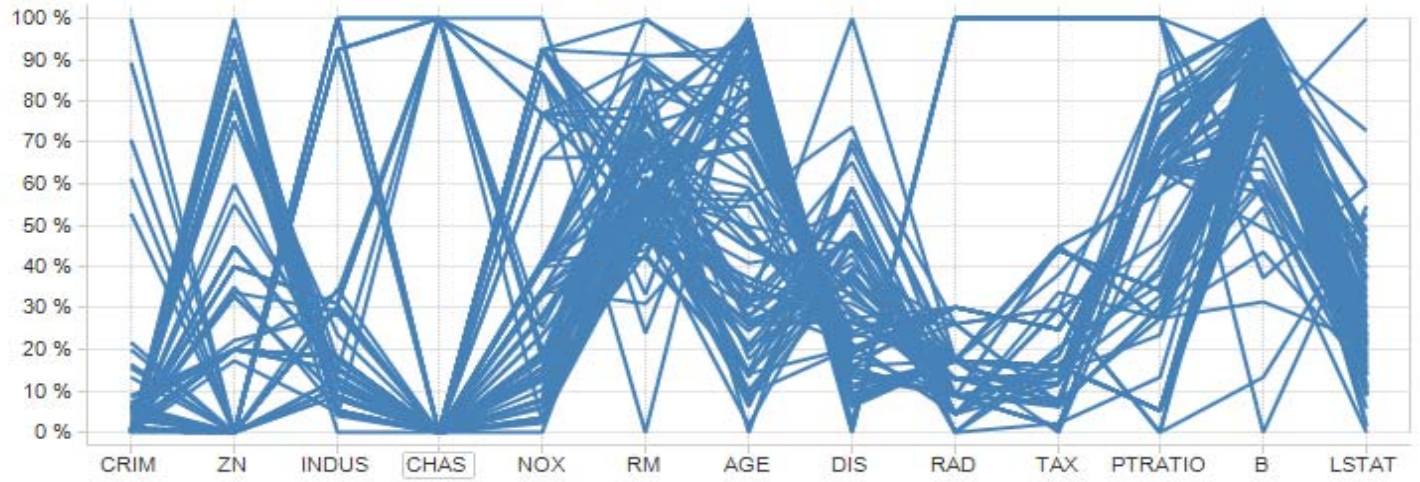
- Uncrowds the data by allowing more markers to be seen 把聚集在一起的记号打散，便于看到更多的记号。

Without jittering (for comparison)

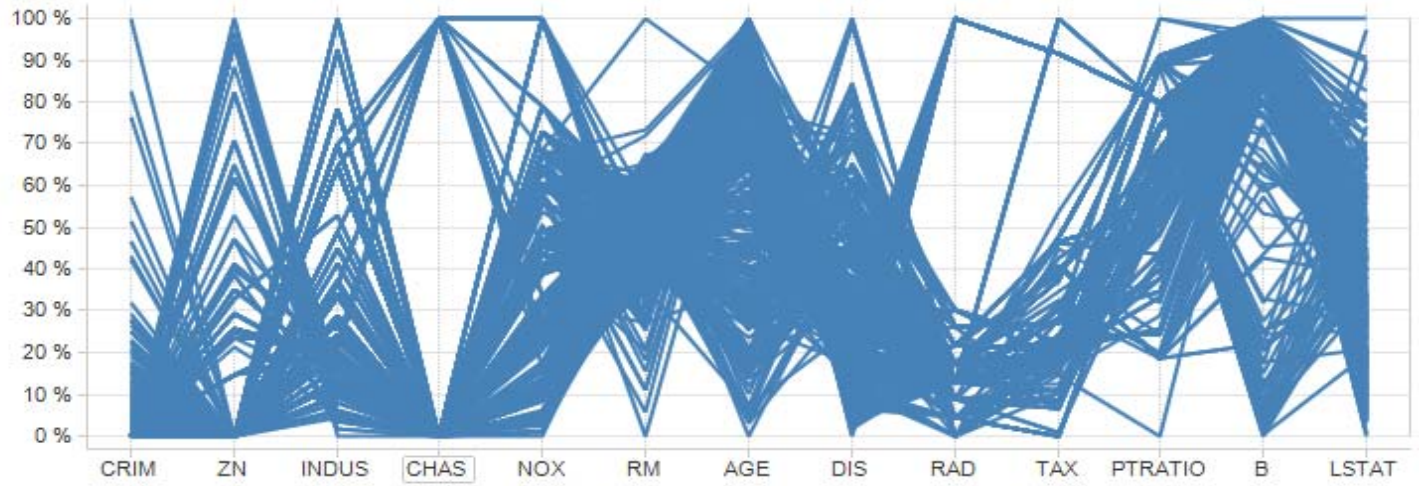


Parallel Coordinate Plot (Boston Housing) 平行坐标图

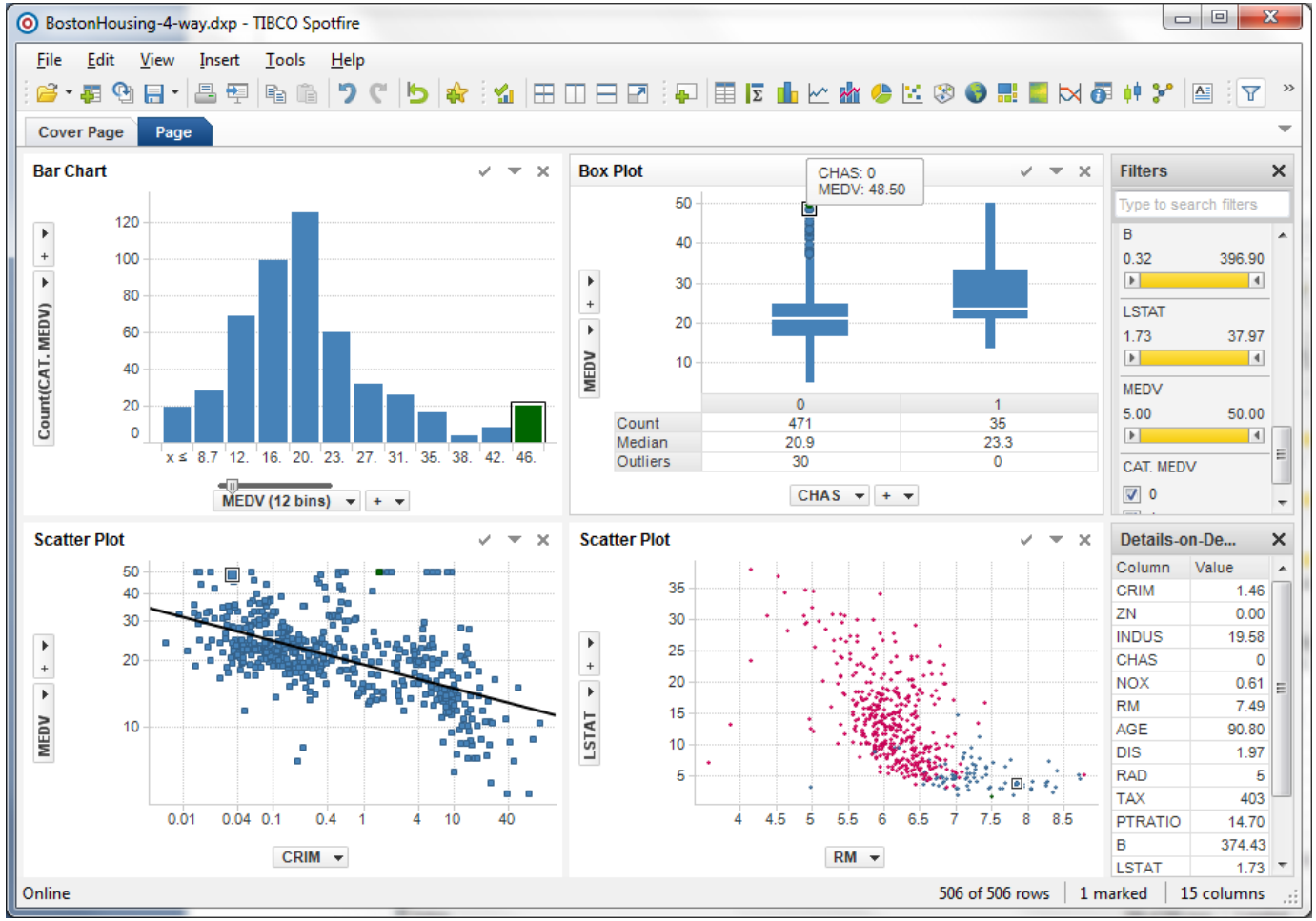
CATMEDV =1



CATMEDV =0



Linked plots
(same record is highlighted in each plot)

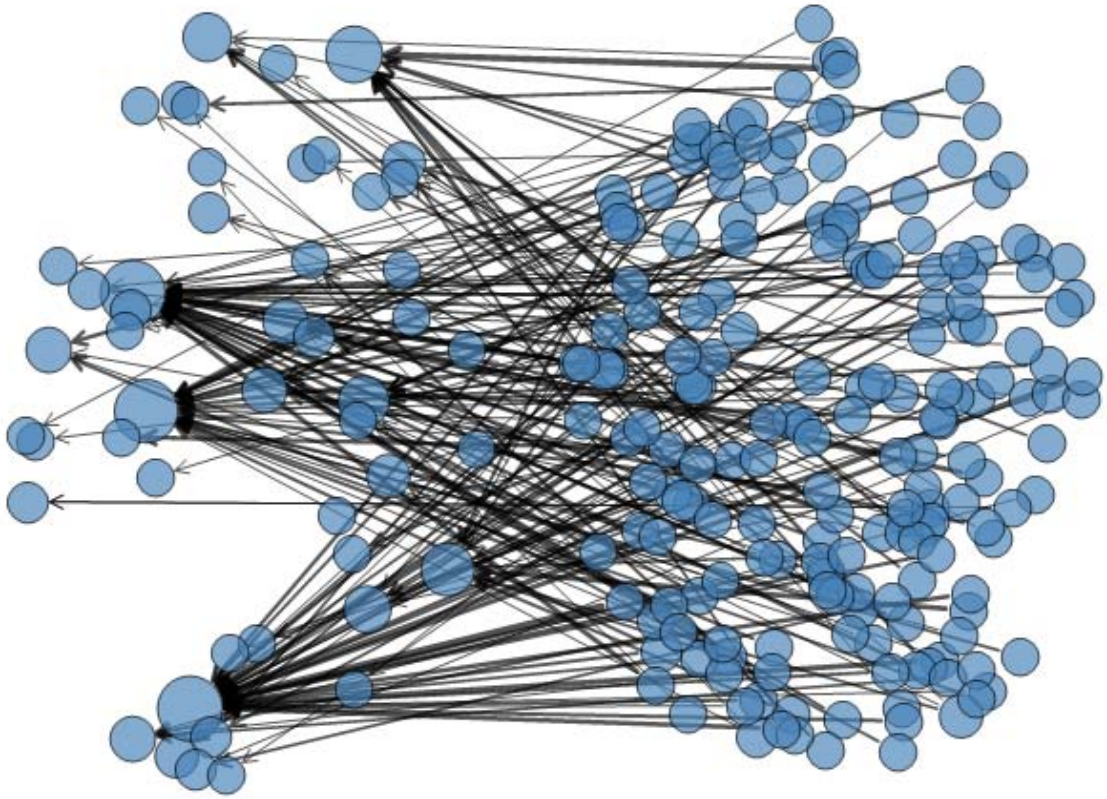


Network Graph 网络图 – eBay Auctions (sellers on left, buyers on right)

Circle size = # of transactions for the node

Line width =# of auctions for the buyer-seller pair

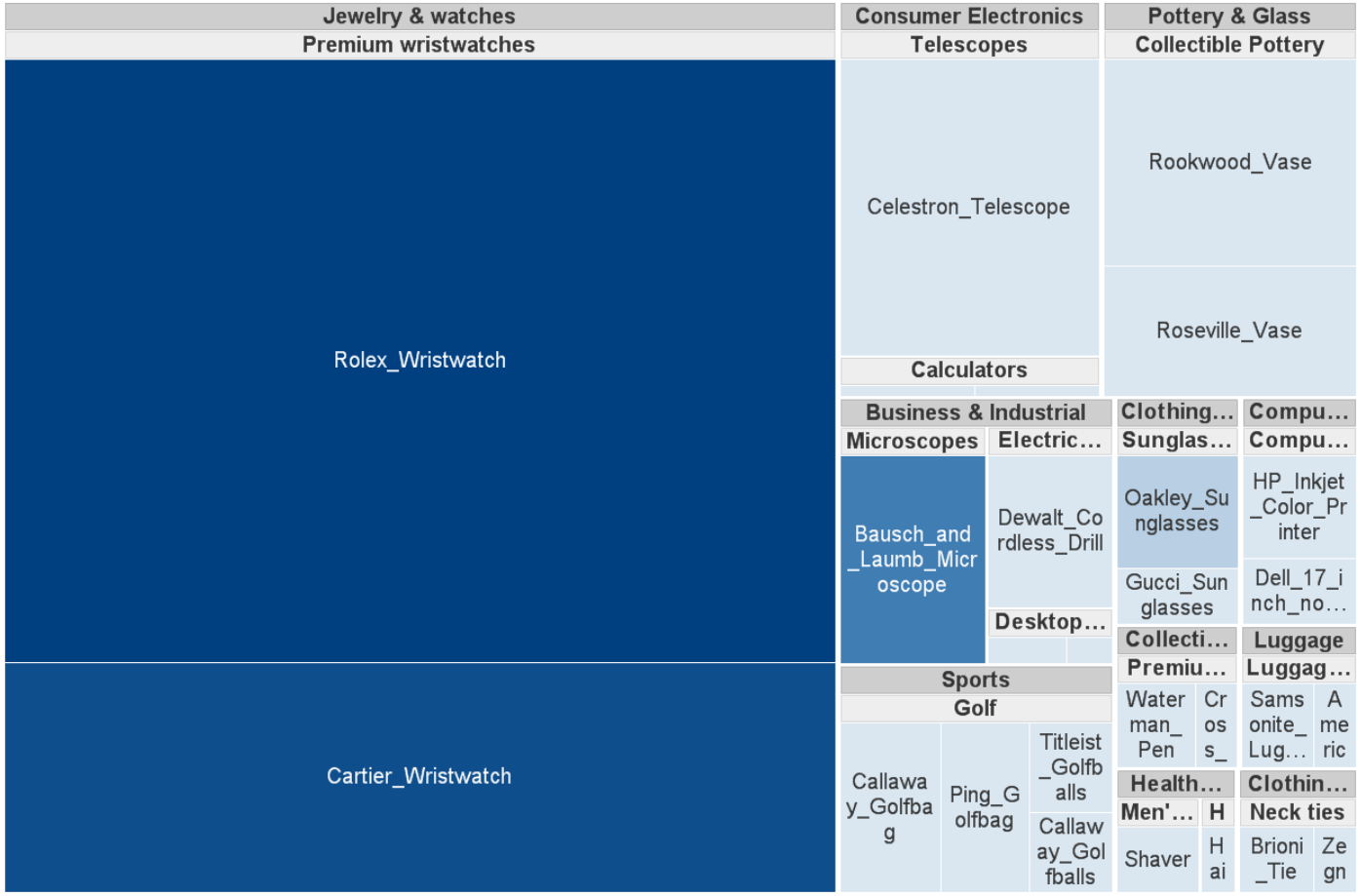
Arrows point from buyer to seller



Treemap – eBay Auctions (Hierarchical eBay data: Category> sub-category> Brand)

Rectangle size = average closing price (=item value)

Color = % sellers with negative feedback (darker=more)



Map Chart

(Comparing countries' well-being with GDP)

Darker = higher value

Well-Being Score

GDP