Chapter 3 – Data Visualization

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

Graphs for Data Exploration

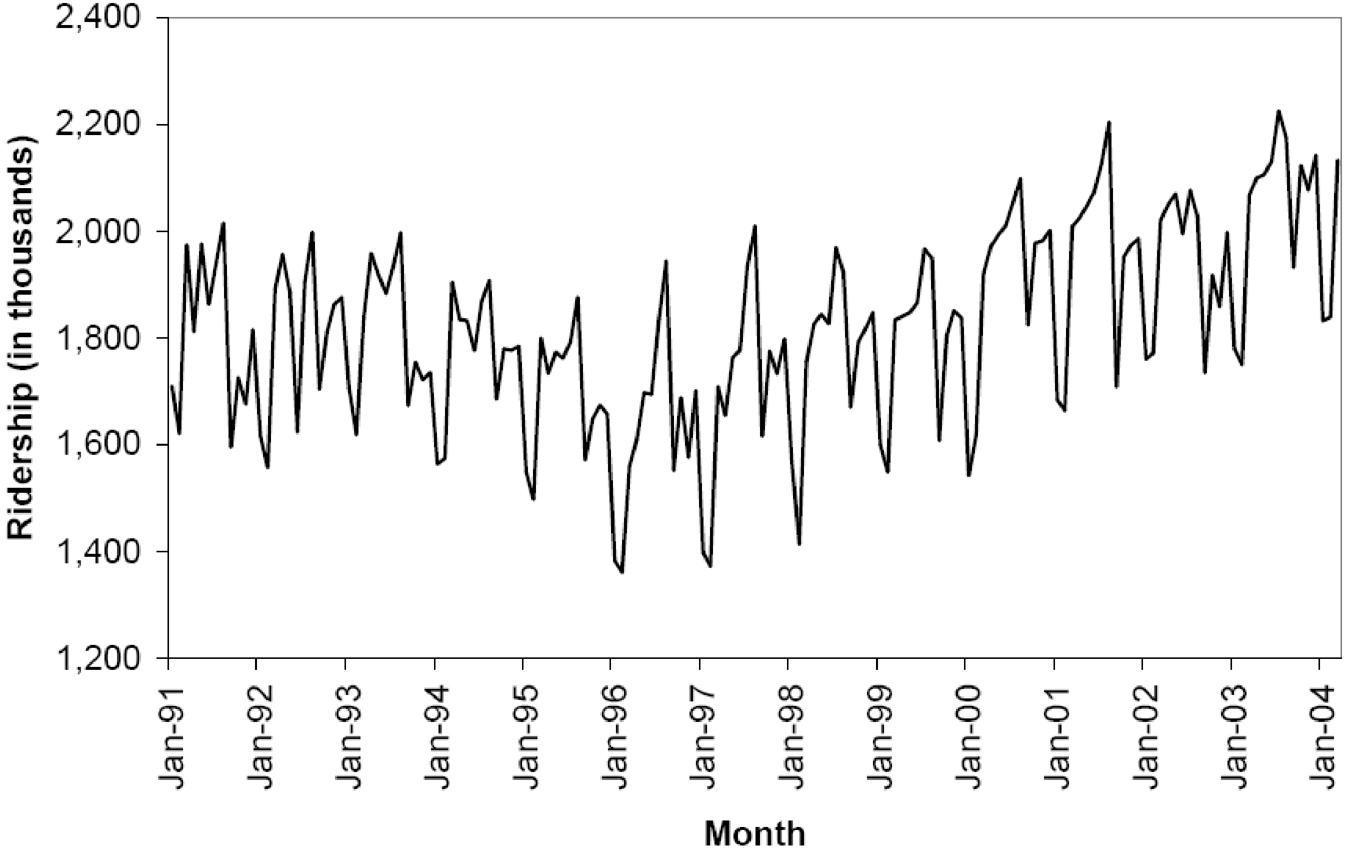
Basic Plots

- Line Graphs
- Bar Charts
- Scatterplots

Distribution Plots

- Boxplots
- Histograms

Line Graph for Time Series

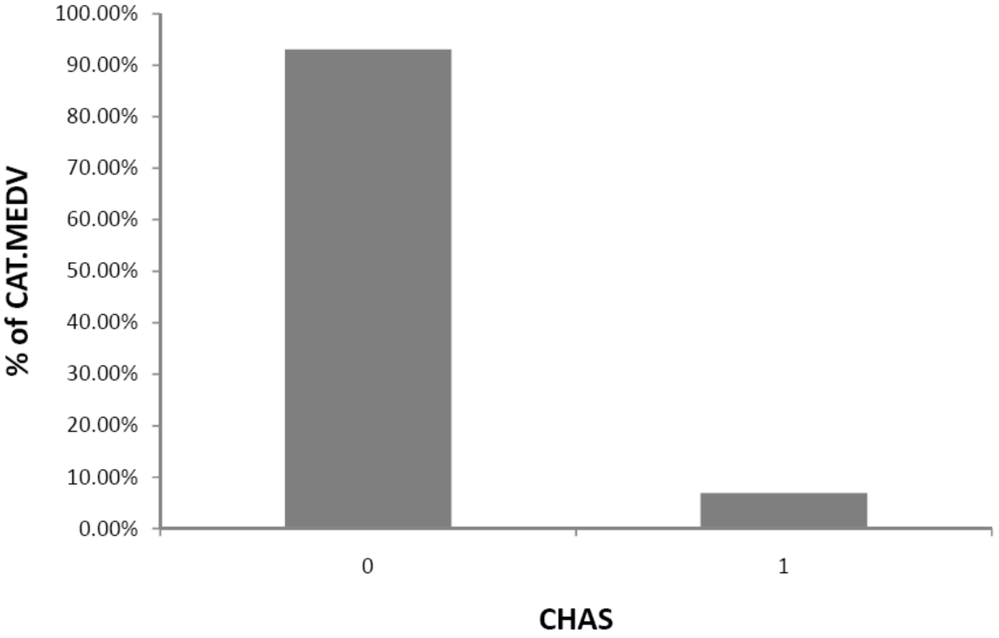


Bar Chart for Categorical Variable

95% of tracts do not border Charles River

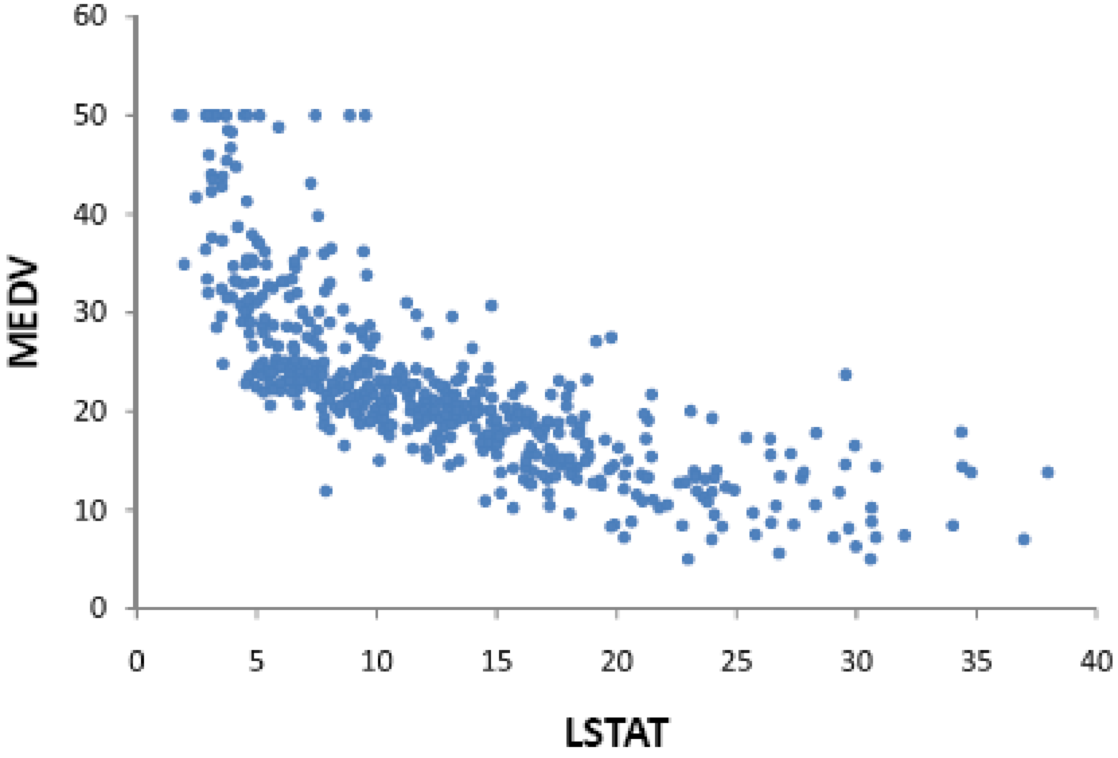
Excel can confuse:

y-axis is actually "% of records that have a value for CATMEDV" (i.e., "% of all records")



Scatterplot

Displays relationship between two numerical variables



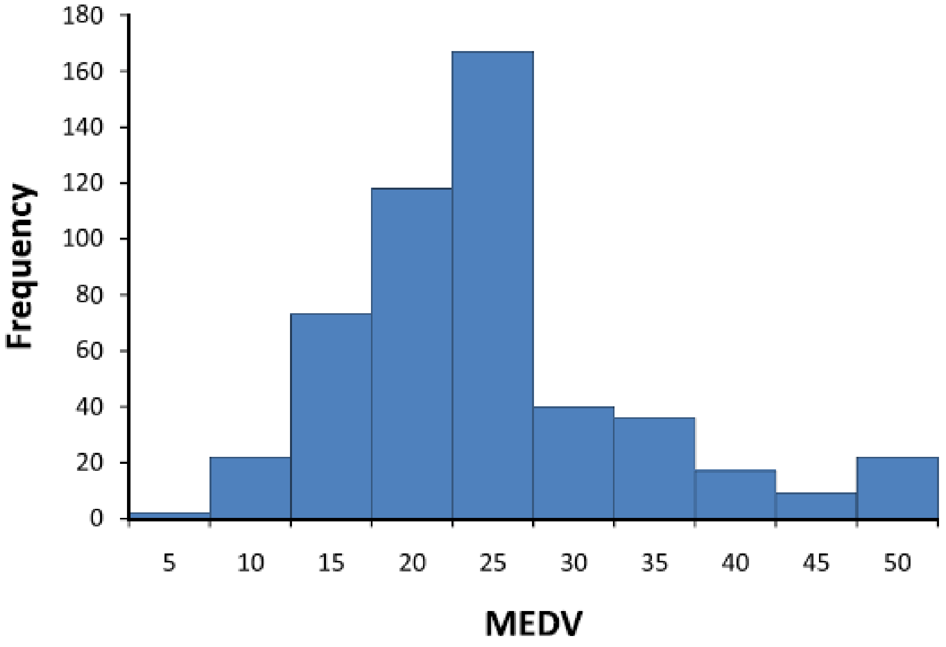
Distribution Plots

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

Histograms

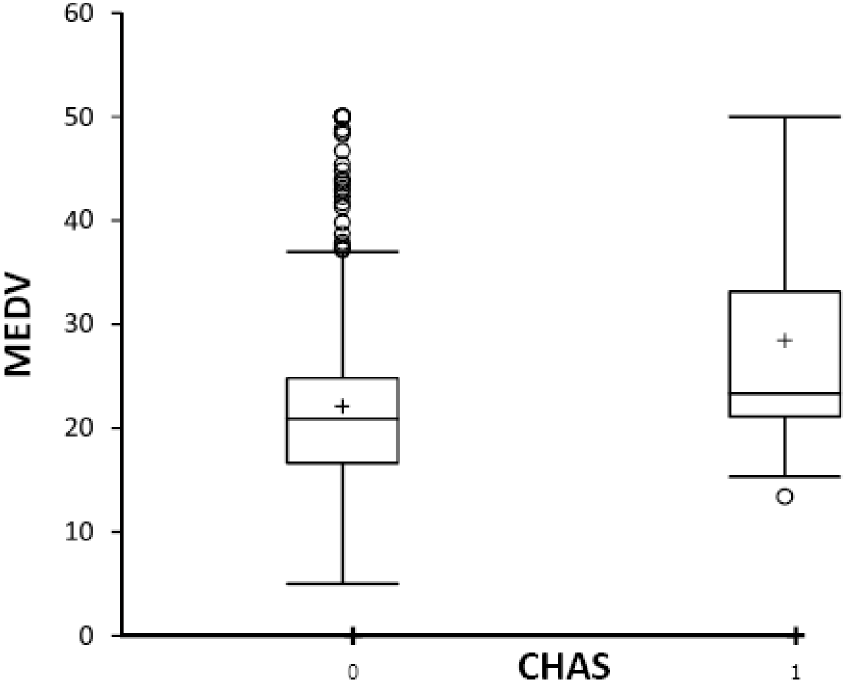
Boston Housing example:

Histogram shows the distribution of the outcome variable (median house value)



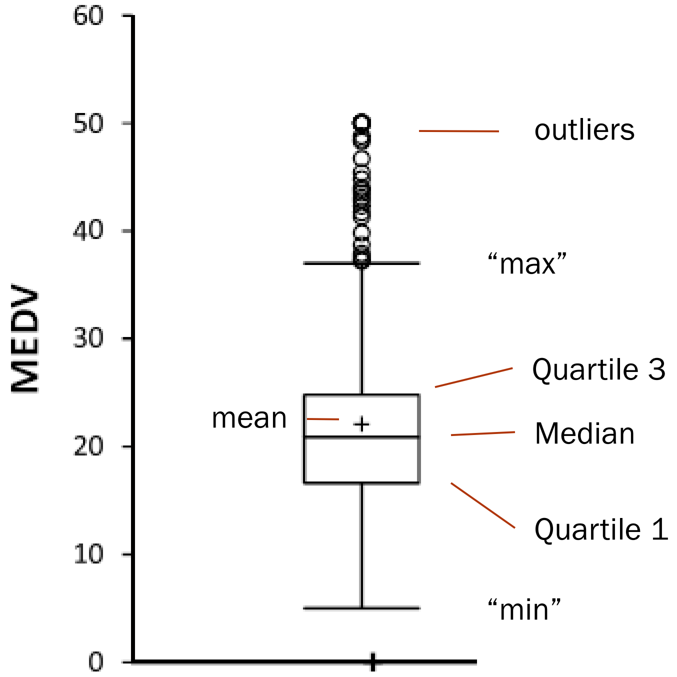
Boxplots

Side-by-side boxplots are useful for comparing subgroups



Boston Housing Example: Display distribution of outcome variable (MEDV) for neighborhoods on Charles river (1) and not on Charles river (0)

Box Plot



- 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

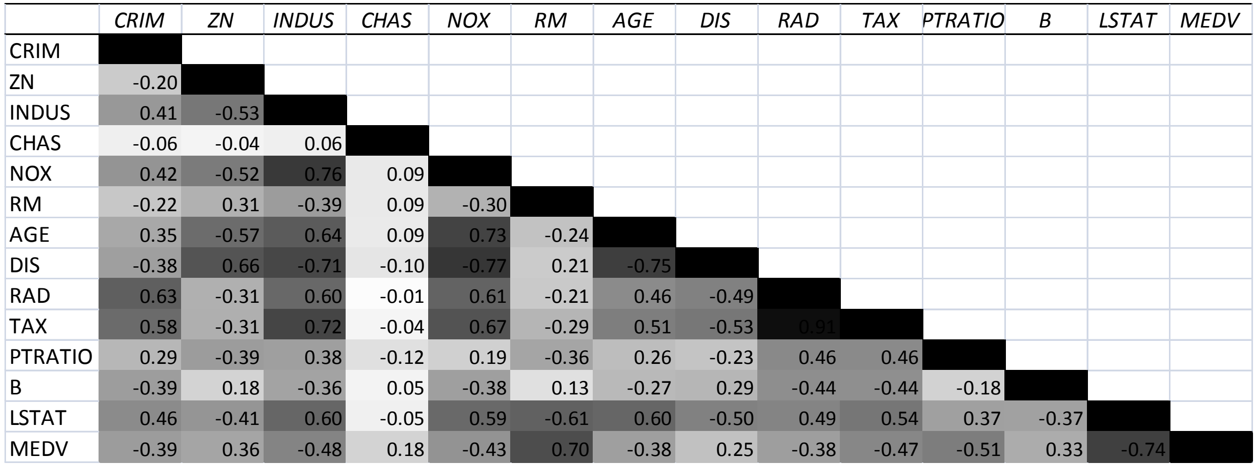
Heat Maps

Color conveys information

In data mining, used to visualize

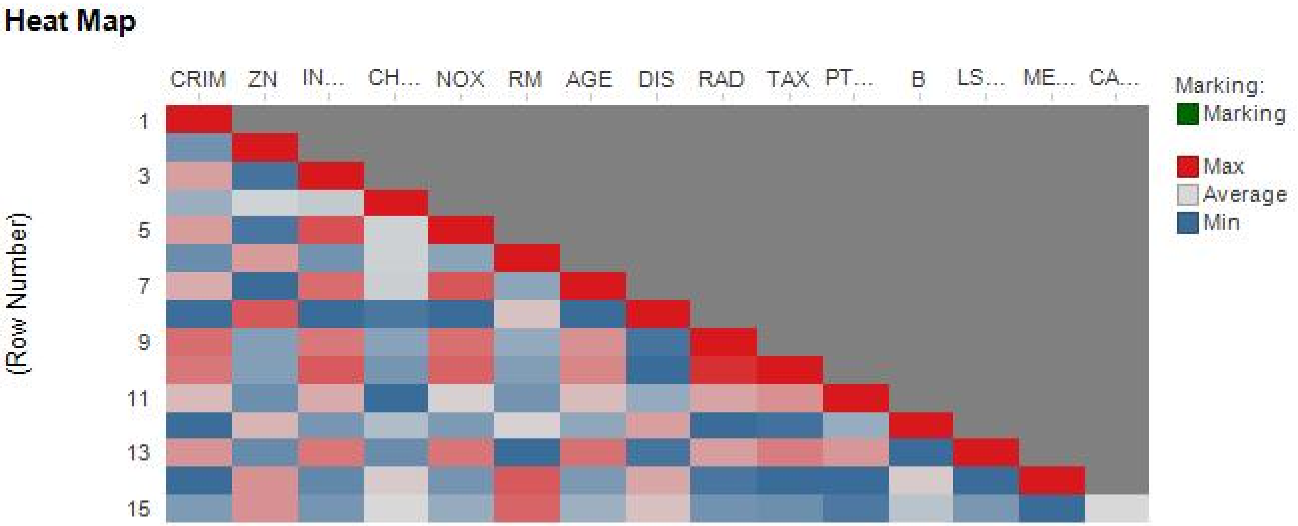
- Correlations
- Missing Data

Heatmap to highlight correlations (Boston Housing)



	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

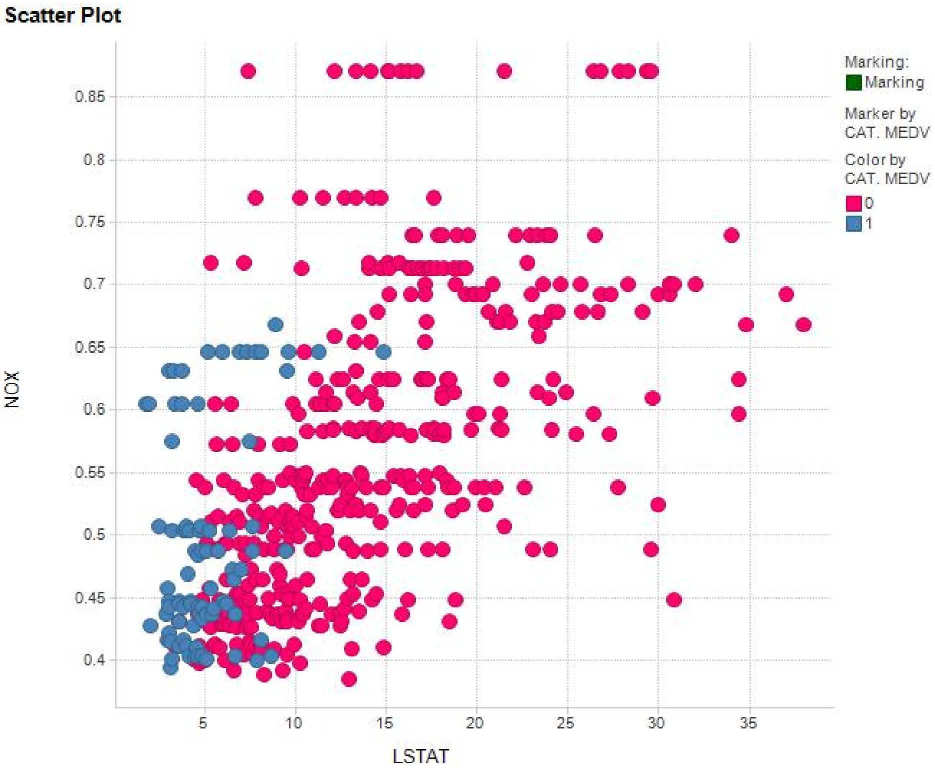
Scatterplot with color added

Boston Housing

NOX vs. LSTAT

Red = low median value

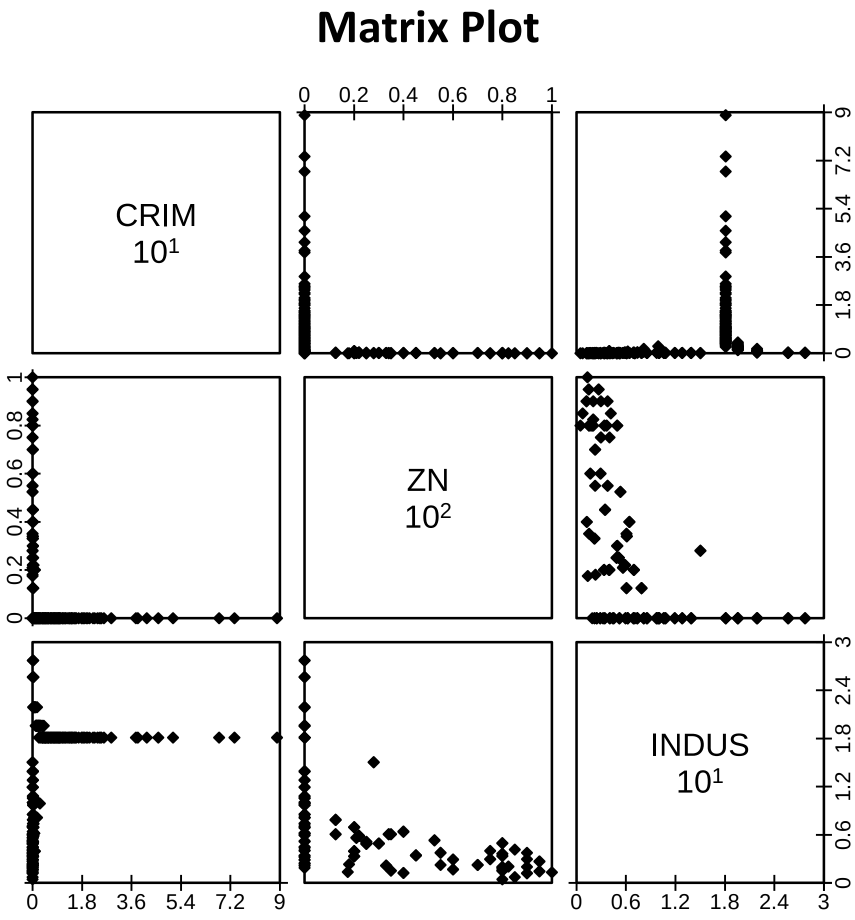
Blue = high median value



Matrix Plot

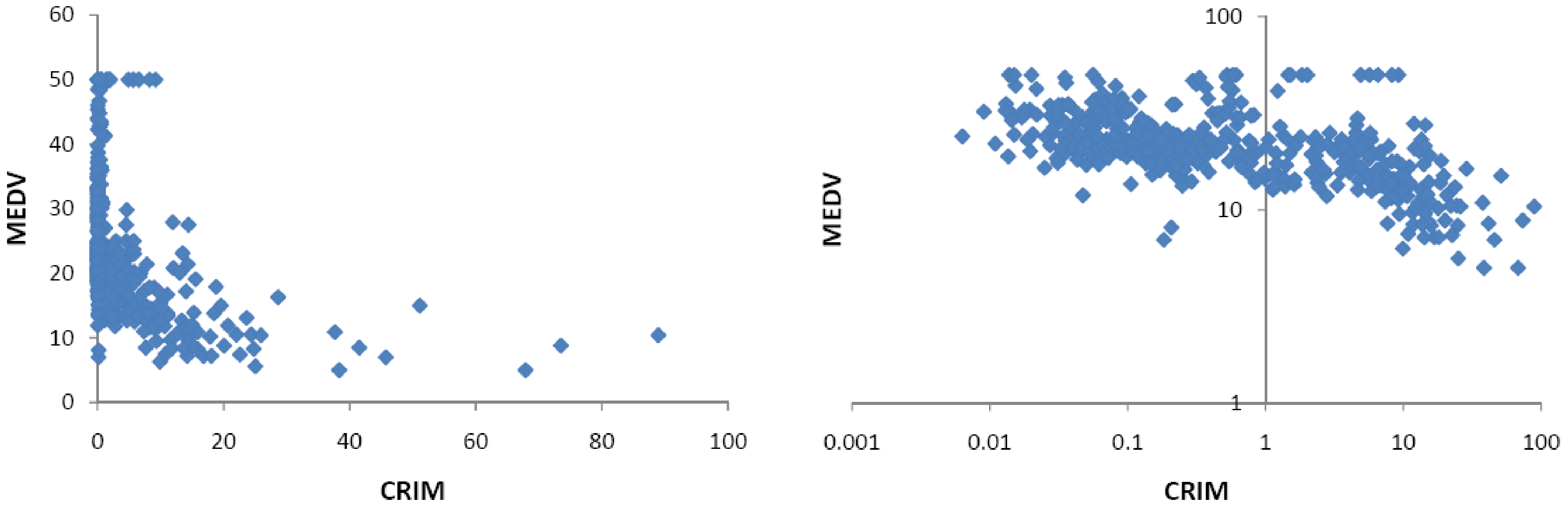
Shows scatterplots for variable pairs

Example: scatterplots for 3 Boston Housing variables



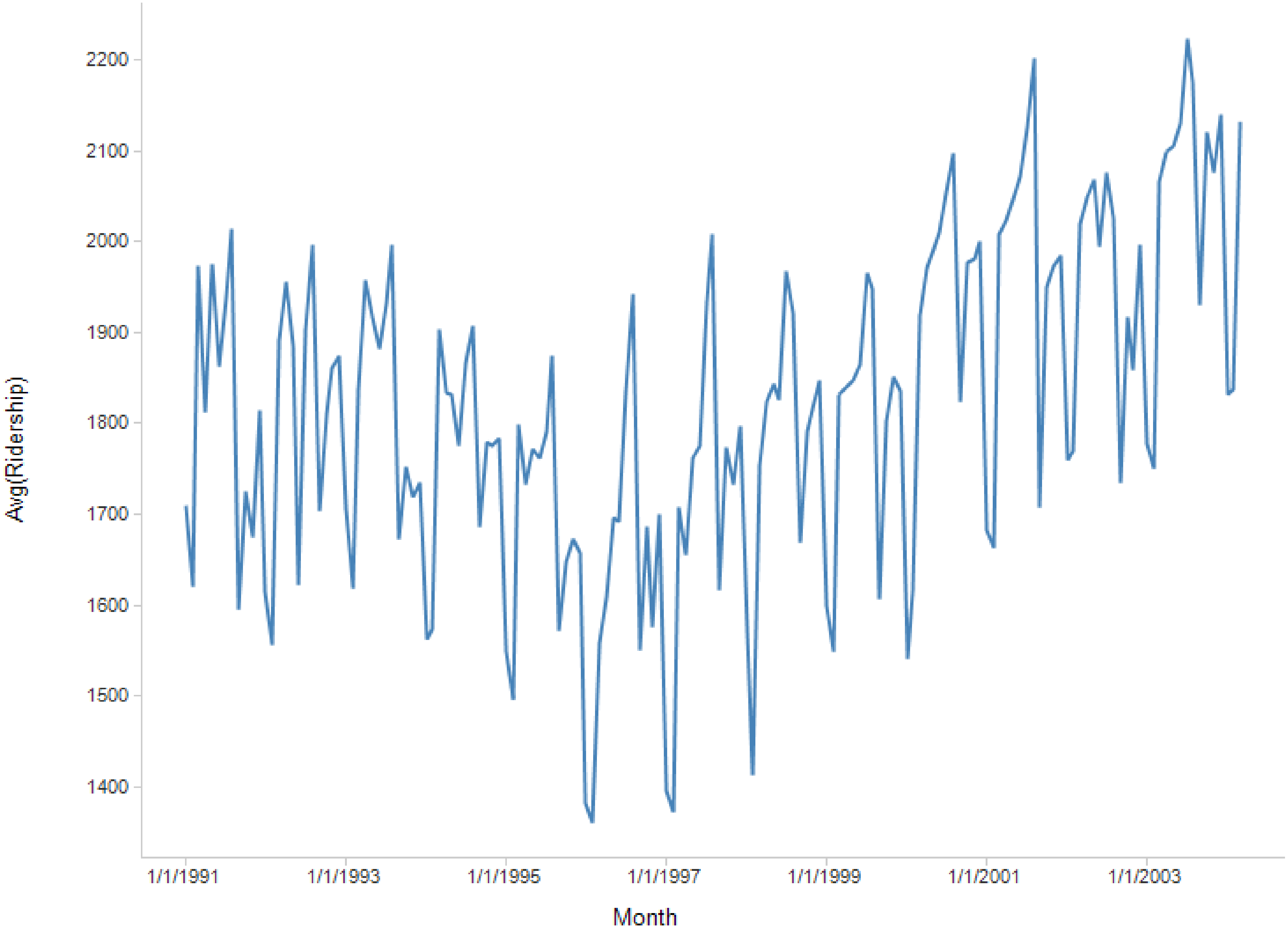
Rescaling to log scale (on right)

“uncrowds” the data

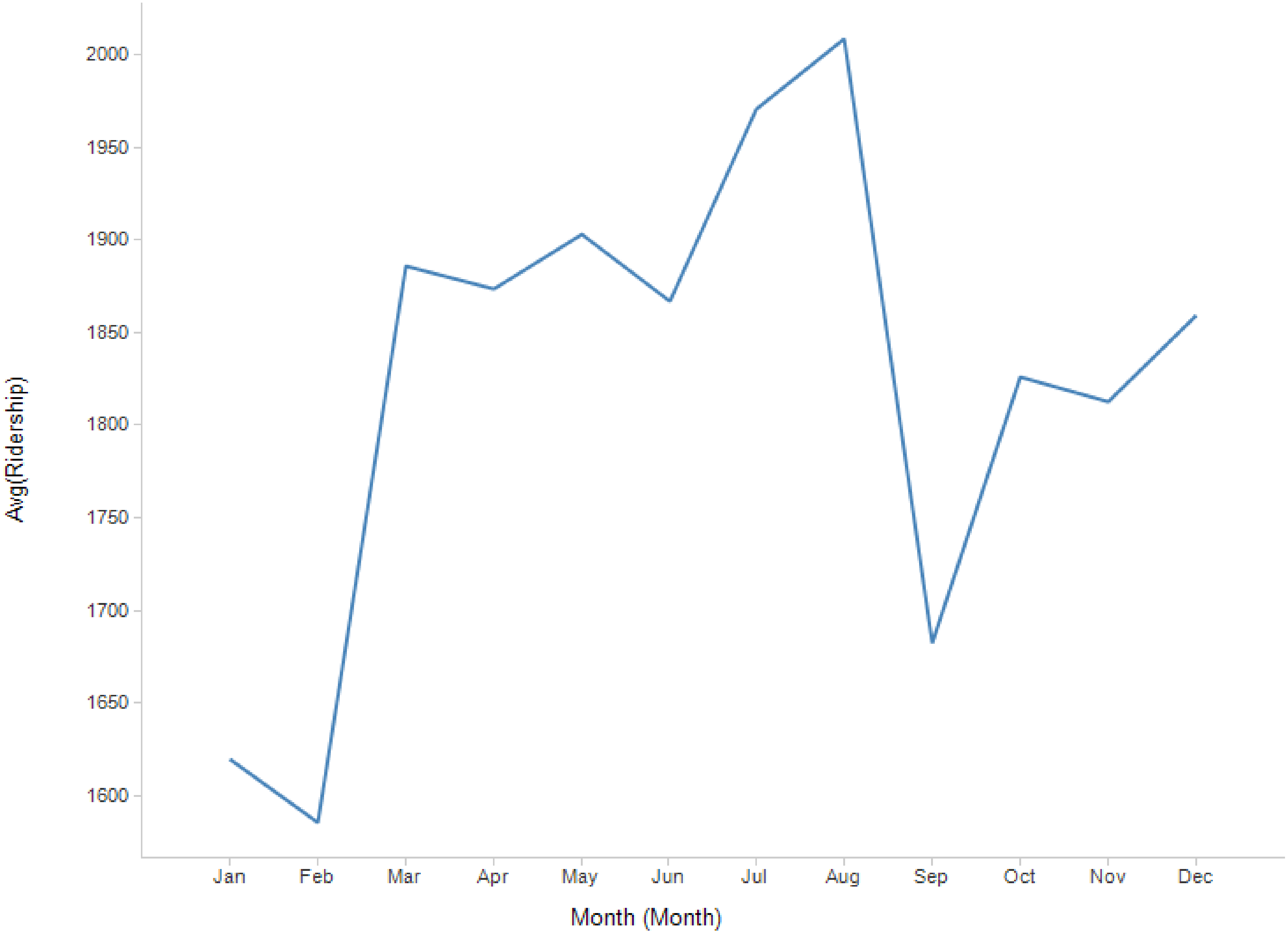


Aggregation

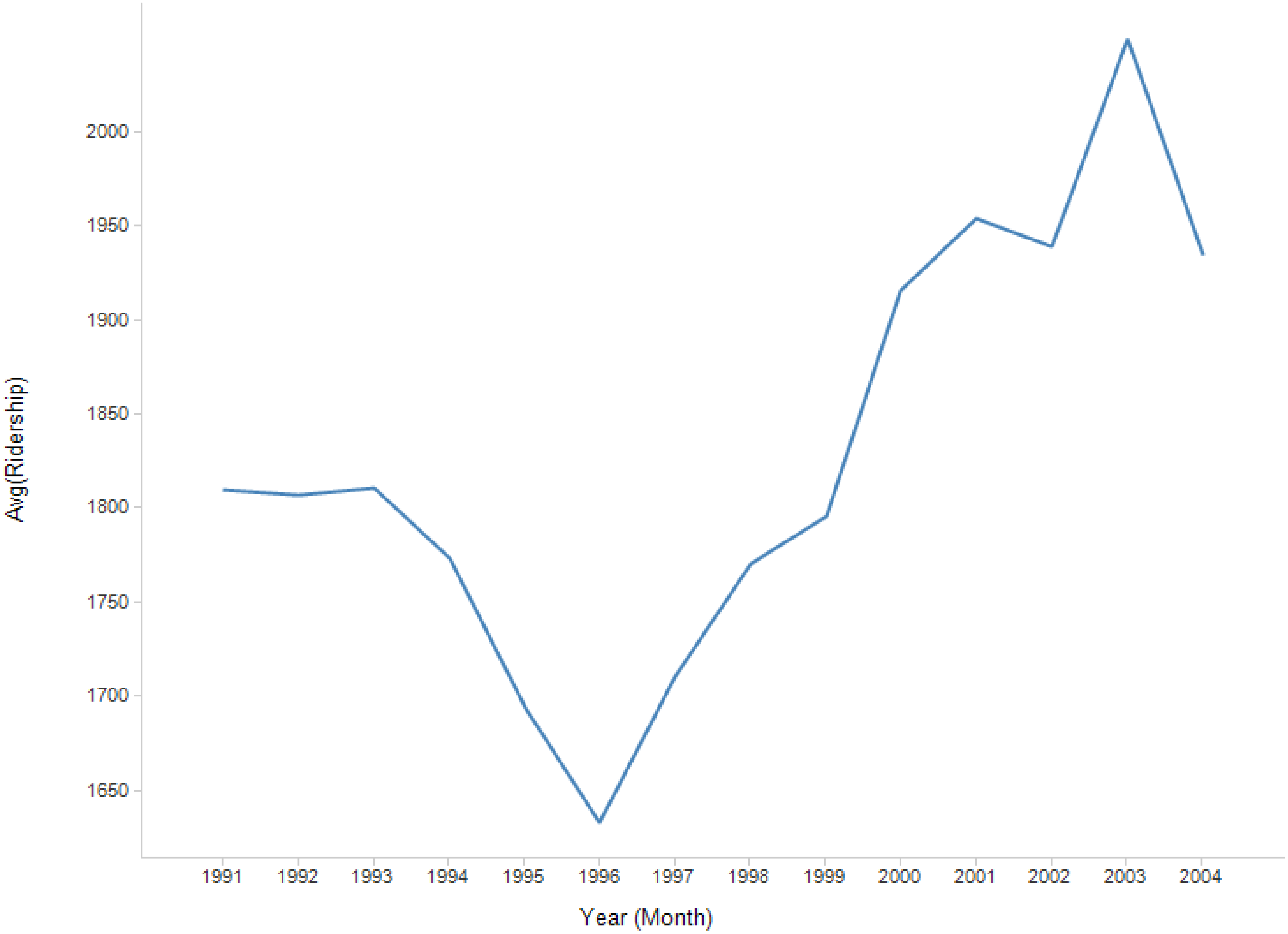
Amtrak Ridership – Monthly Data



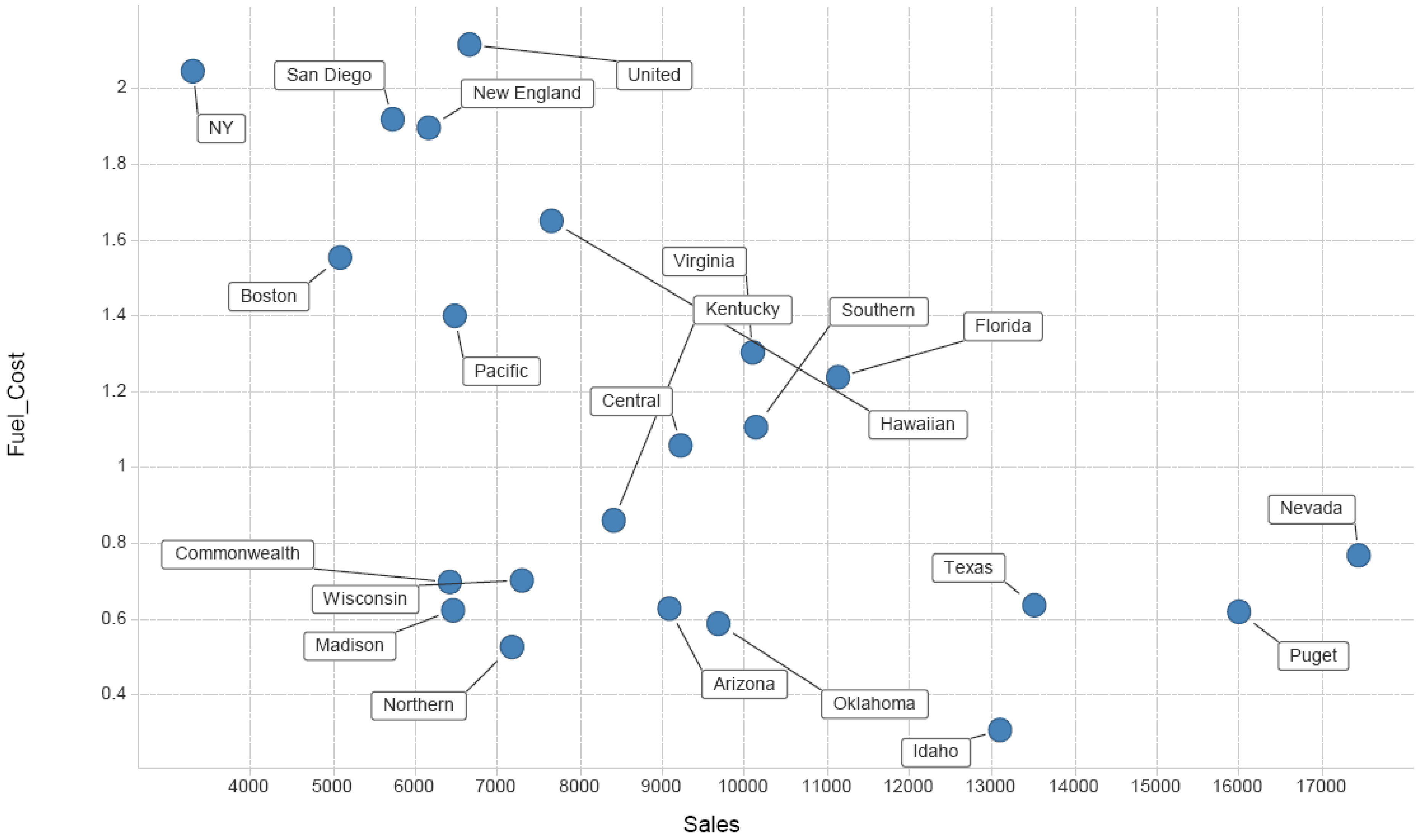
Aggregation – Monthly Average



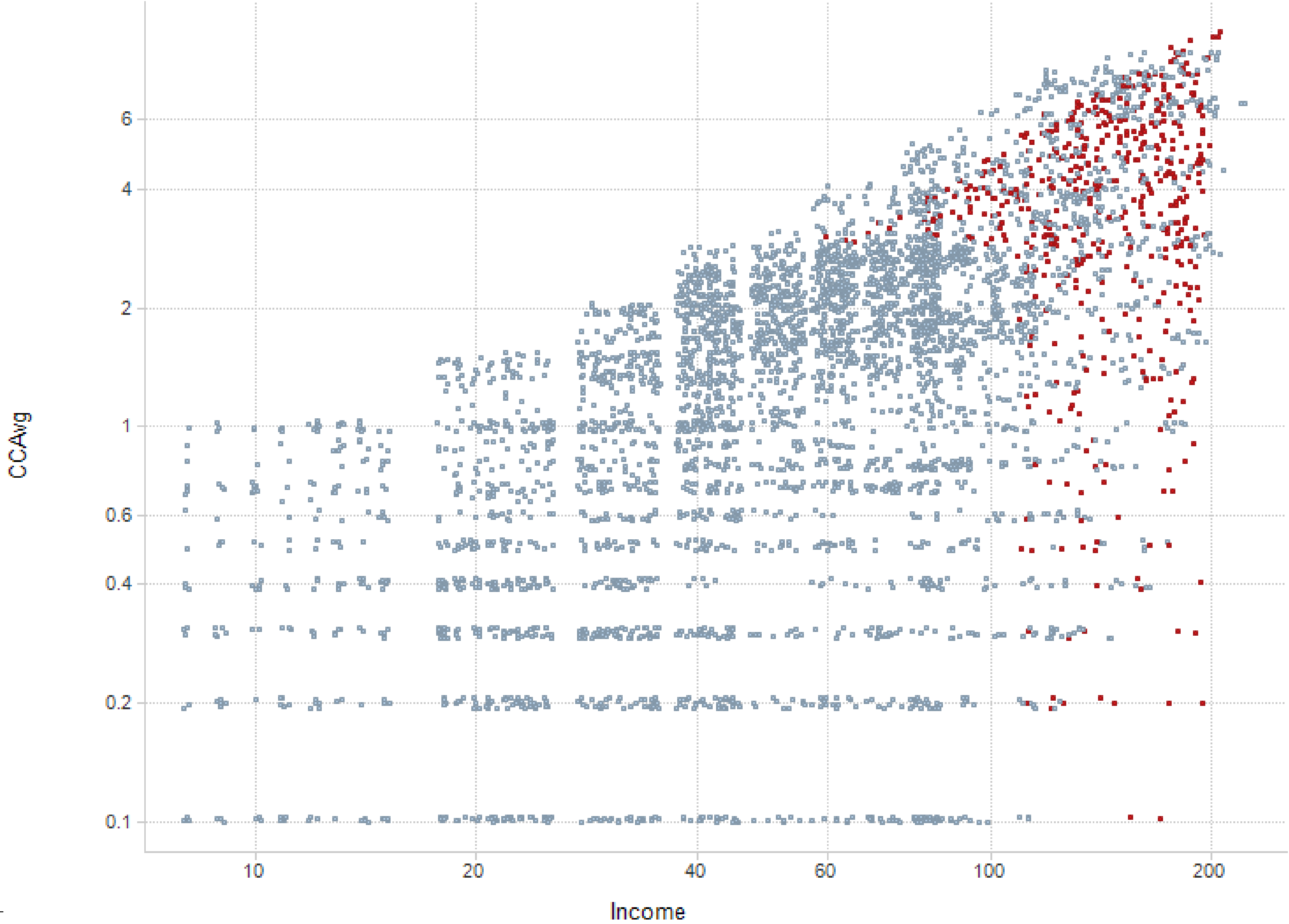
Aggregation – Yearly Average



Scatter Plot with Labels (Utilities)



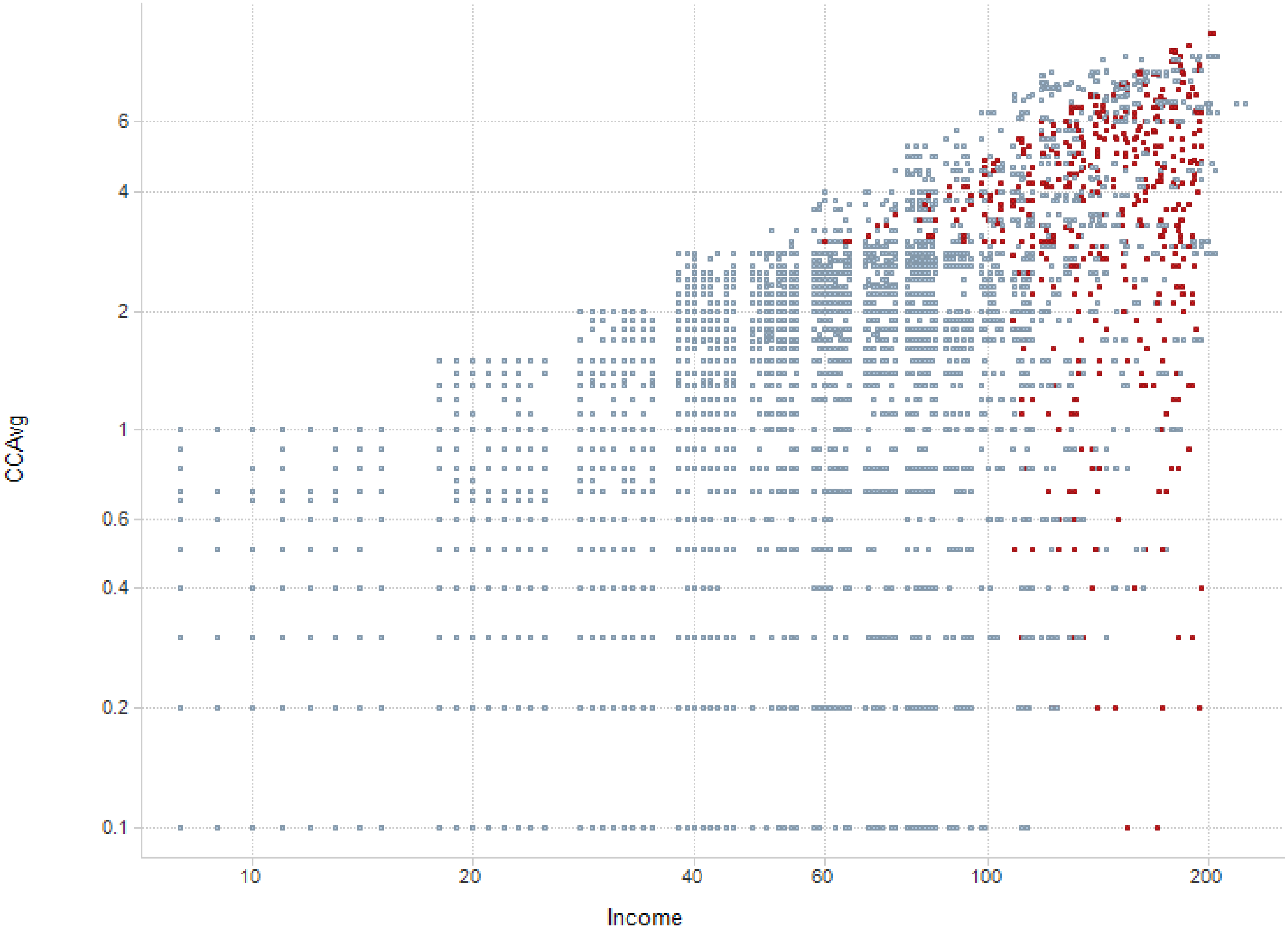
Scaling: Smaller markers, jittering, color contrast (Universal Bank; red = accept loan)



Jittering

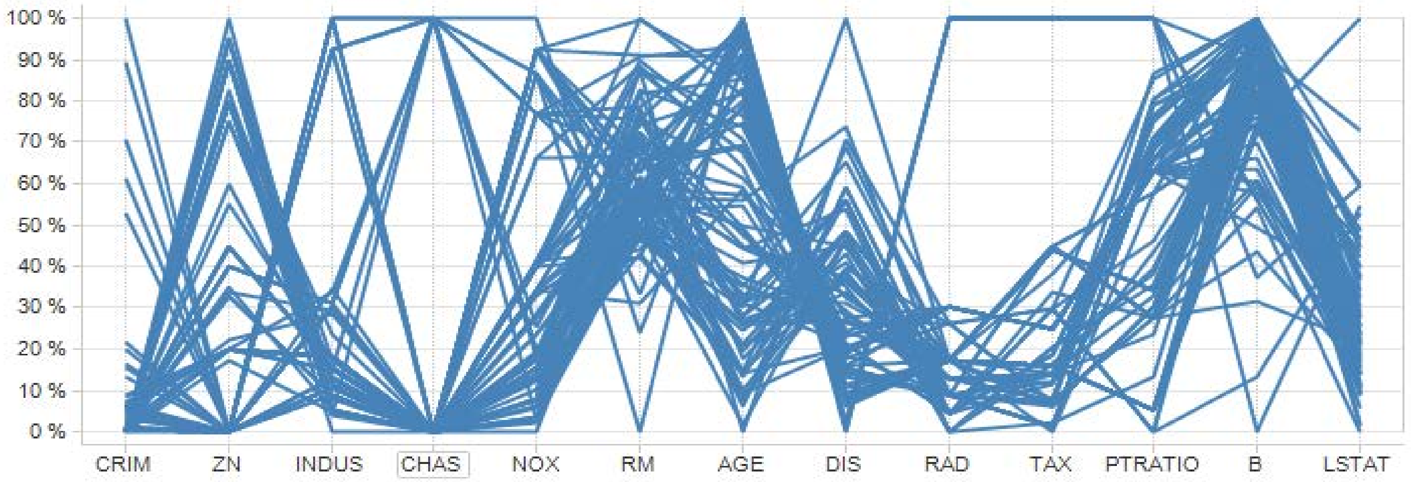
- Moving markers by a small random amount
- Uncrowds the data by allowing more markers to be seen

Without jittering (for comparison)

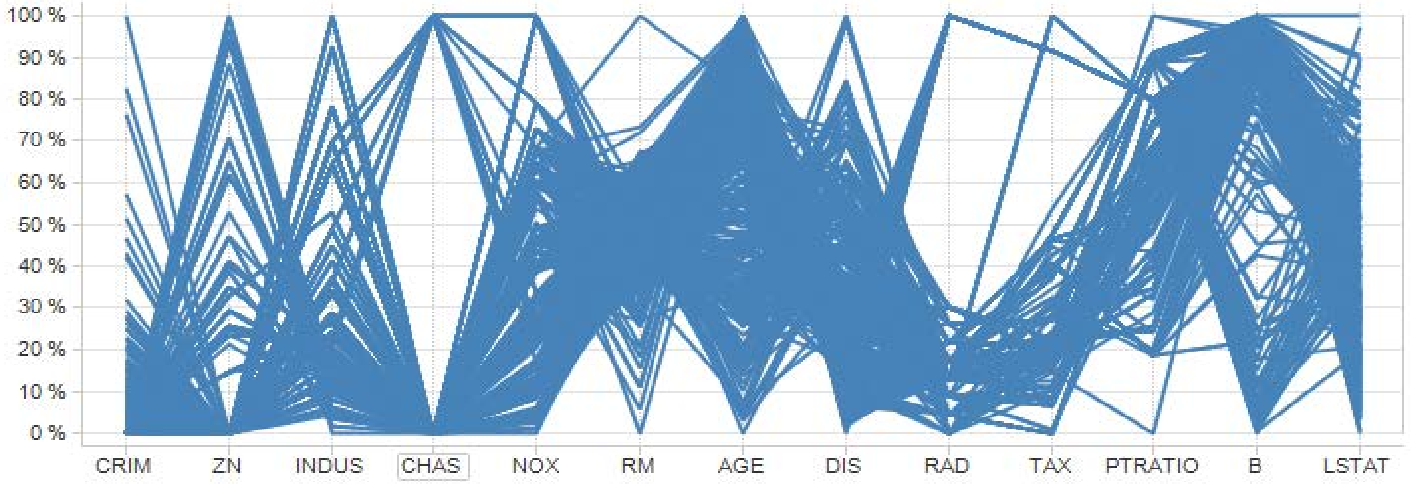


Parallel Coordinate Plot (Boston Housing)

CATMEDV =1

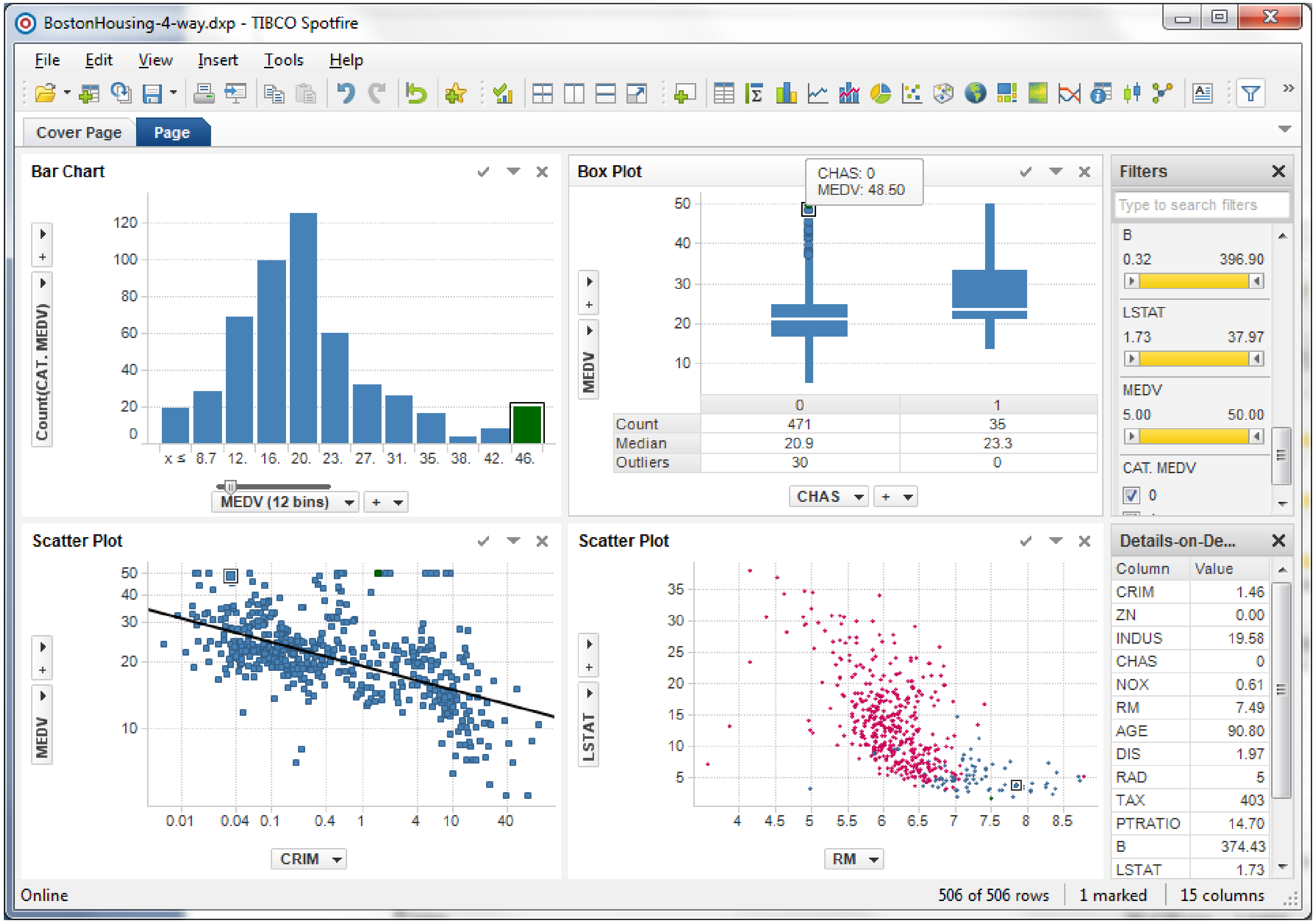


CATMEDV =0



Linked plots

(same record is highlighted in each plot)



	0	1
Count	471	35
Median	20.9	23.3
Outliers	30	0

Filters

B
0.32 396.90

LSTAT
1.73 37.97

MEDV
5.00 50.00

CAT. MEDV
☑ 0

Column	Value
CRIM	1.46
ZN	0.00
INDUS	19.58
CHAS	0
NOX	0.61
RM	7.49
AGE	90.80
DIS	1.97
RAD	5
TAX	403
PTRATIO	14.70
B	374.43
LSTAT	1.73

Online | 506 of 506 rows | 1 marked | 15 columns

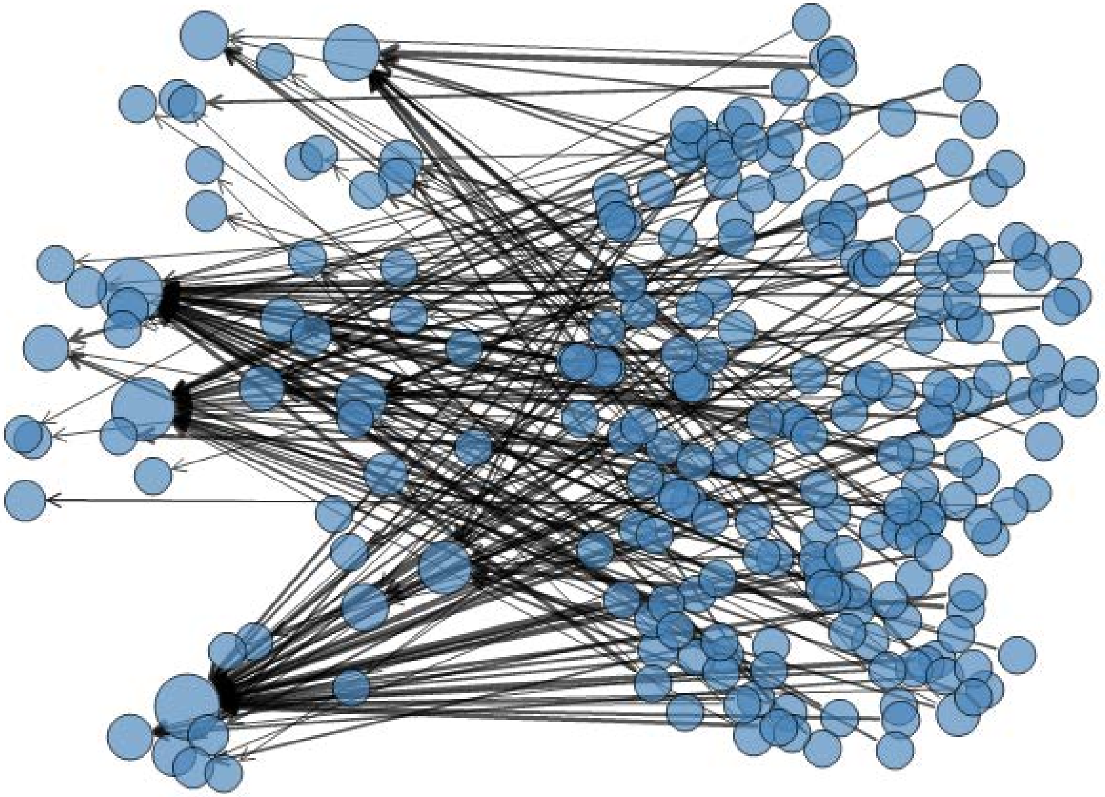
Network Graph – eBay Auctions

(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

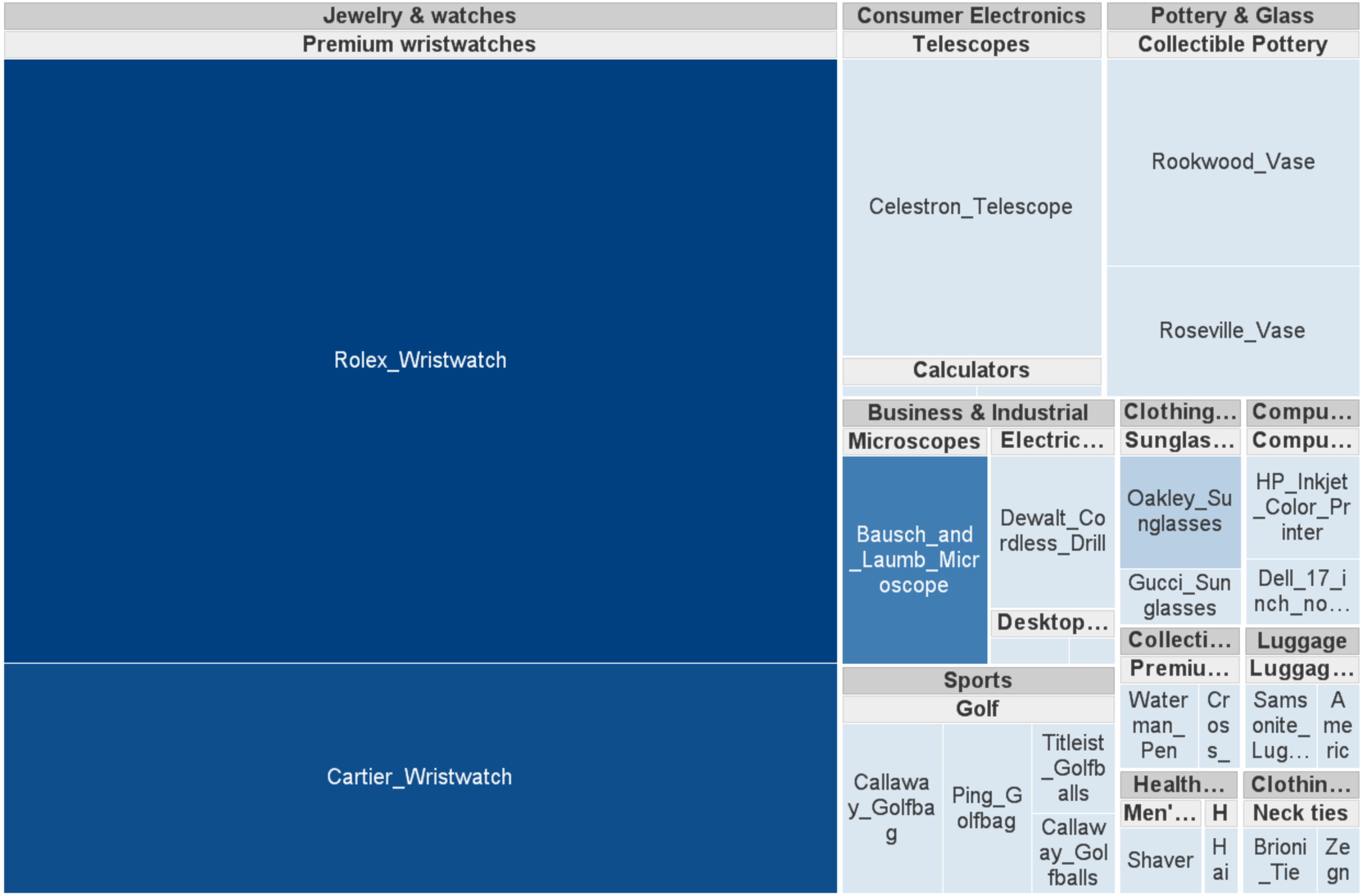


Treemap – eBay Auctions

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

Rectangle size = average closing price (=item value)

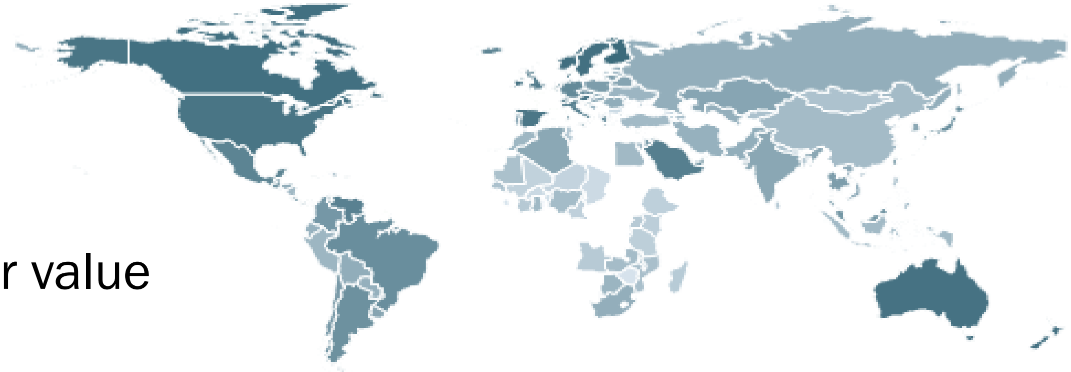
Color = % sellers with negative feedback (darker=more)



Map Chart

(Comparing countries' well-being with GDP)

Well-Being Score



Darker = higher value

GDP