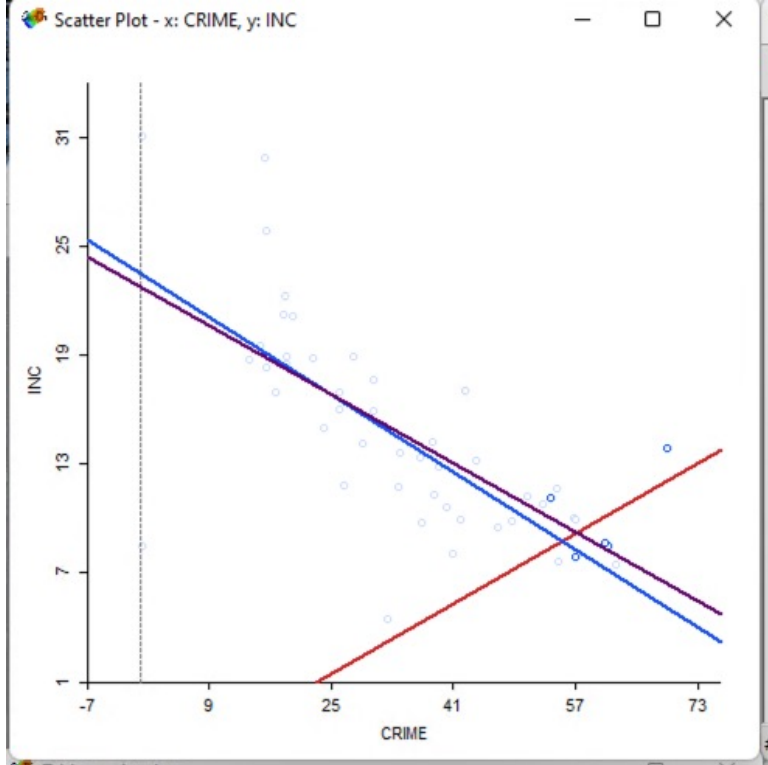
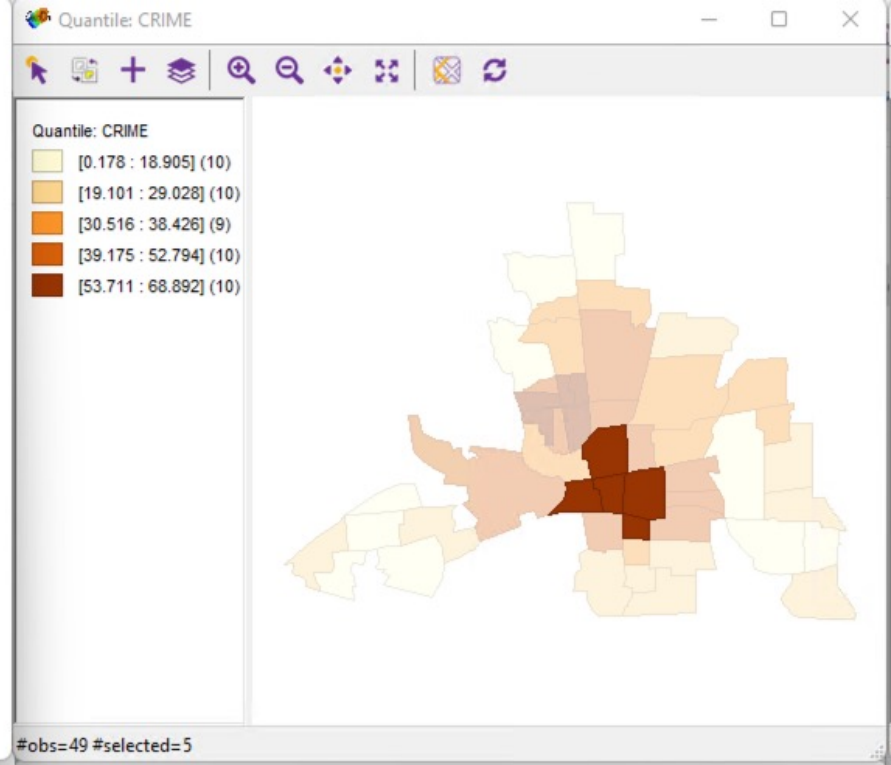
GEOG 413/613

LECTURE 9

Multivariate Exploratory Data Analysis

- Exploratory Data Analysis
  - the initial investigations on data
    - discover patterns
    - Reduce dimensions
    - identify anomalies/outliers
    - test hypothesis (e.g. observed vs expected)
    - check assumptions
      - Descriptive statistics
      - Visualization





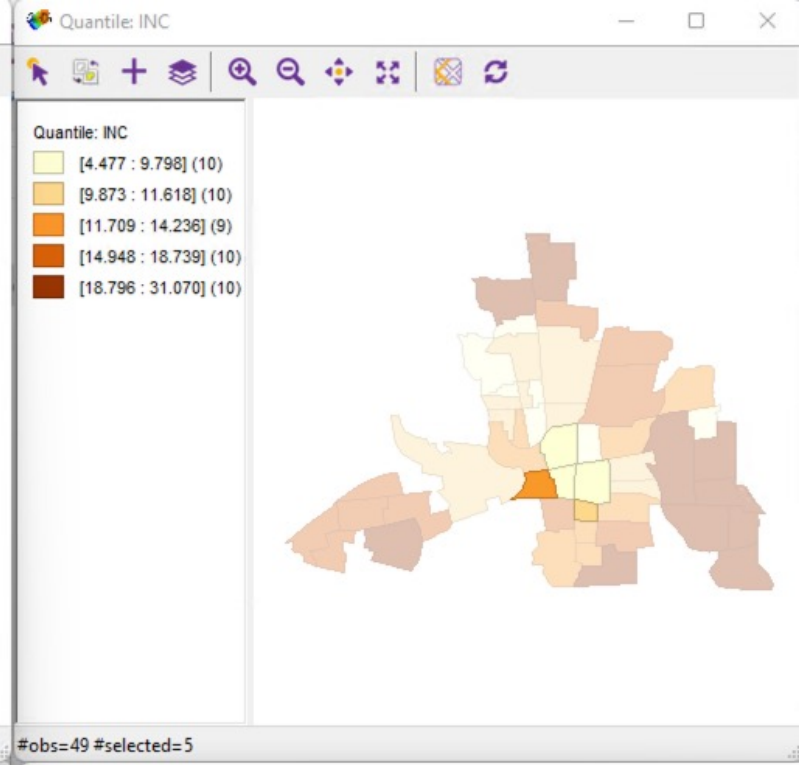
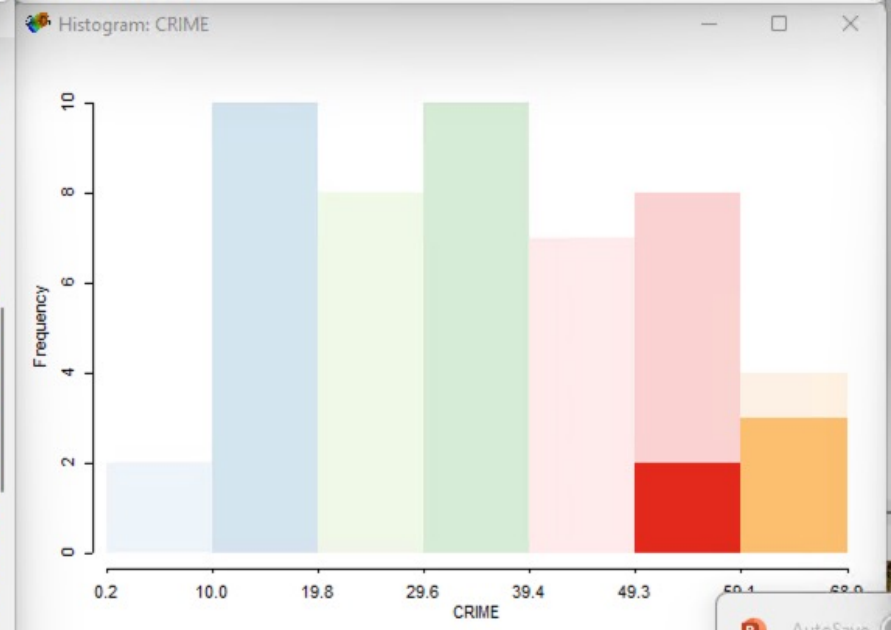
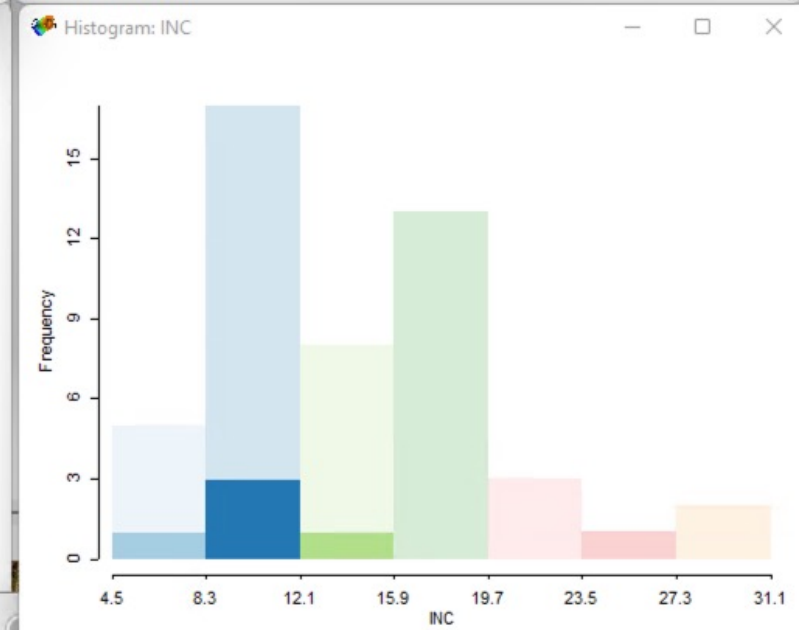


Table - columbus

	NEIG	HOVAL	INC	CRIME	OPEN	PLUMB
24	35	53.200001	14.236000	38.297871	0.626220	18.811075
25	32	17.900000	8.461000	61.299175	0.000000	6.529851
26	20	20.299999	8.085000	40.969742	1.238288	2.534275
27	21	34.099998	10.822000	52.794430	19.368099	1.483516
28	31	22.850000	7.856000	56.919785	0.509305	3.001072
29	33	32.500000	8.681000	60.750446	0.000000	2.645051
30	34	22.500000	13.906000	68.892044	1.638780	15.600624
31	45	31.799999	16.940001	17.677214	3.936443	0.853890
32	13	40.299999	18.941999	19.145592	2.221022	0.255102
33	22	23.600000	9.918000	41.968163	0.000000	1.023891
34	44	28.450001	14.948000	23.974028	3.029087	0.386803
35	23	27.000000	12.814000	39.175053	4.220401	0.633675
36	46	36.299999	18.739000	14.305556	6.773331	0.332349
37	30	43.299999	17.017000	42.445076	4.839273	1.230329
38	24	22.700001	11.107000	53.710938	0.000000	0.800000
39	47	39.599998	18.476999	19.100863	0.000000	0.314663
40	16	61.950001	29.833000	16.241299	6.451310	0.132743
41	14	42.099998	22.207001	18.905146	0.293317	0.247036

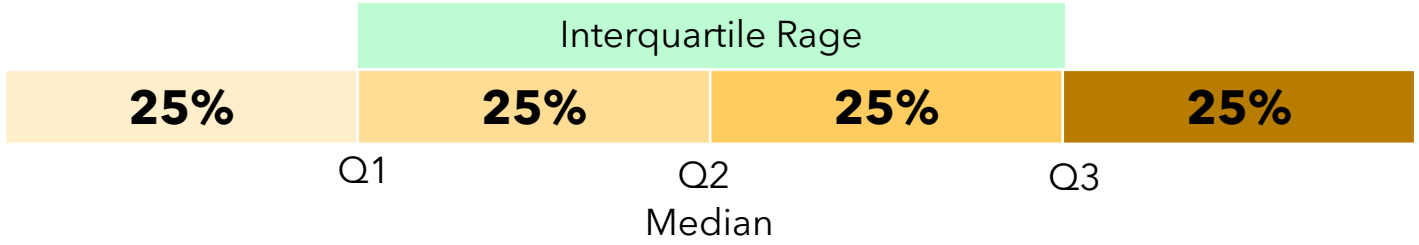
Row=49 #selected=5





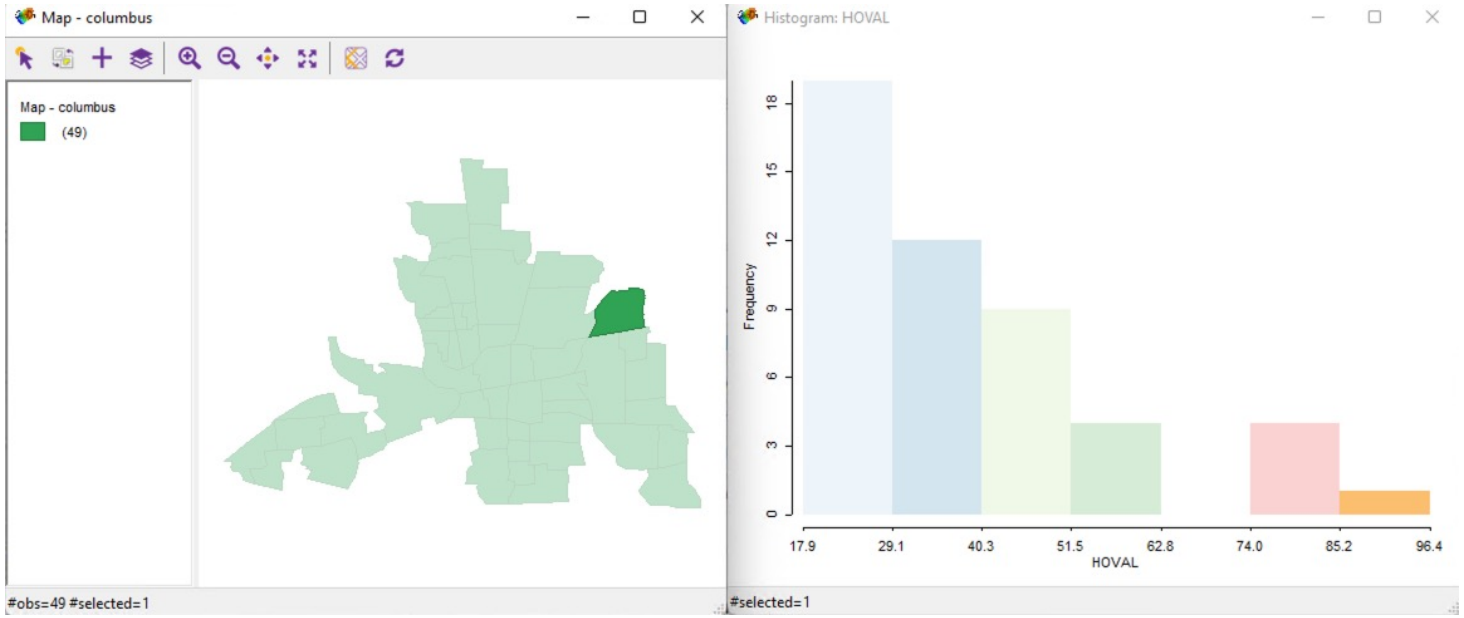
Exploratory Data Analysis

- Outlier detection
  - attribute value(s) are markedly different from others consideration
  - data may be correct
    - may represent the most important items in an investigation (e.g. pollutant source)
  - data may be the erroneous (e.g., measurement error)
    - Warrants removal



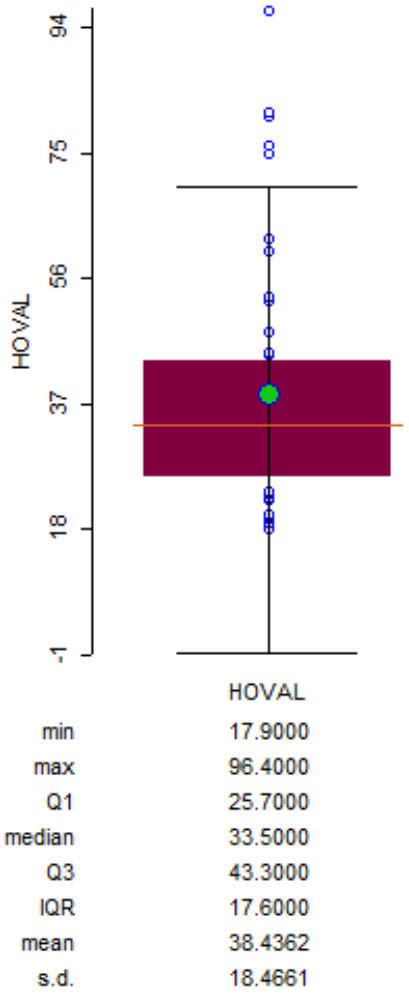
Outlier Detection

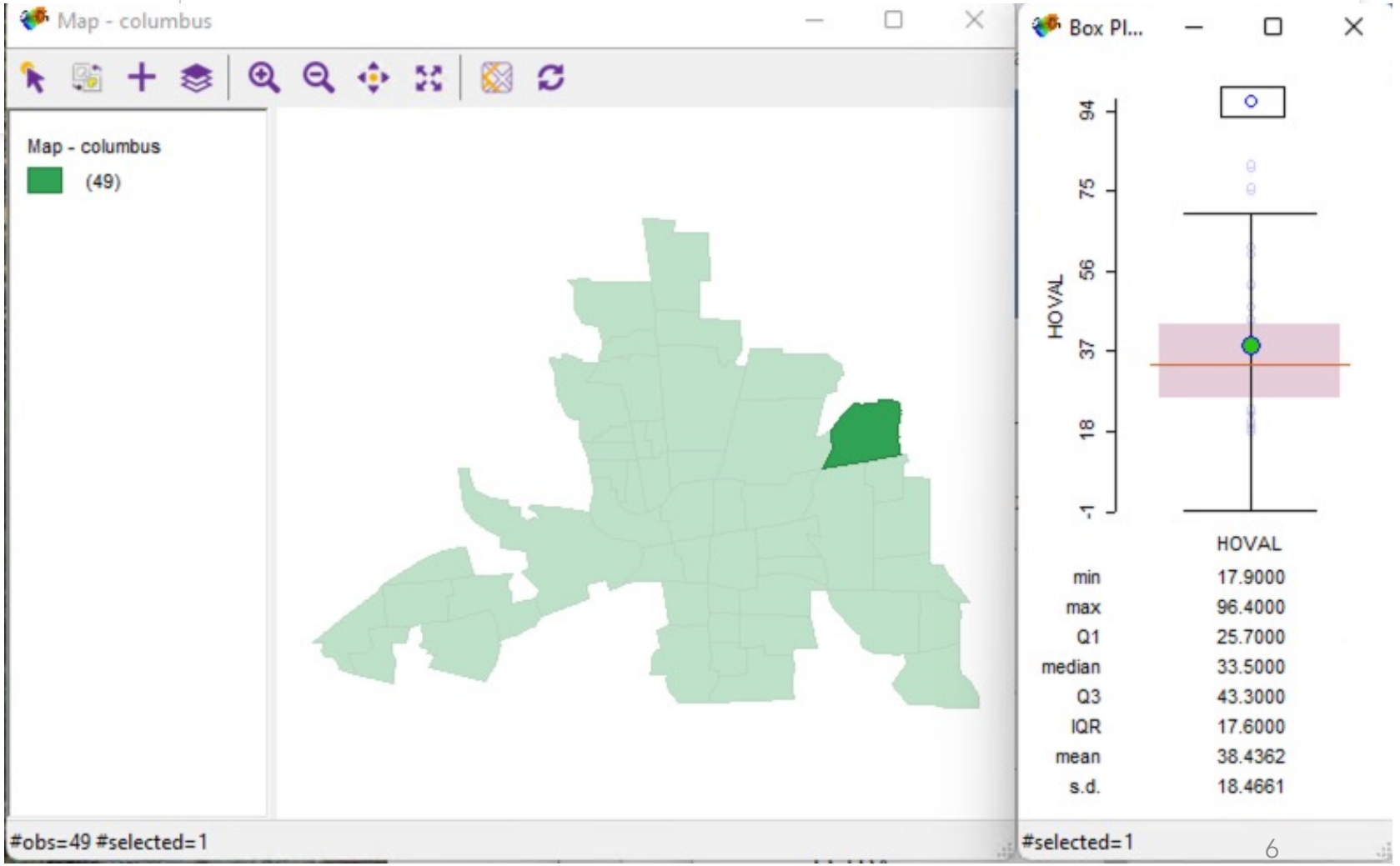
- histograms and mapped histograms
  - Preferable to use a fine class (bin) division, and then identify extreme classes
  - *global outliers* - values that are the limits of the range
  - *local outliers* – relative extremes e.g. markedly different neighbors



Outlier Detection

- Box plots





Box Plots

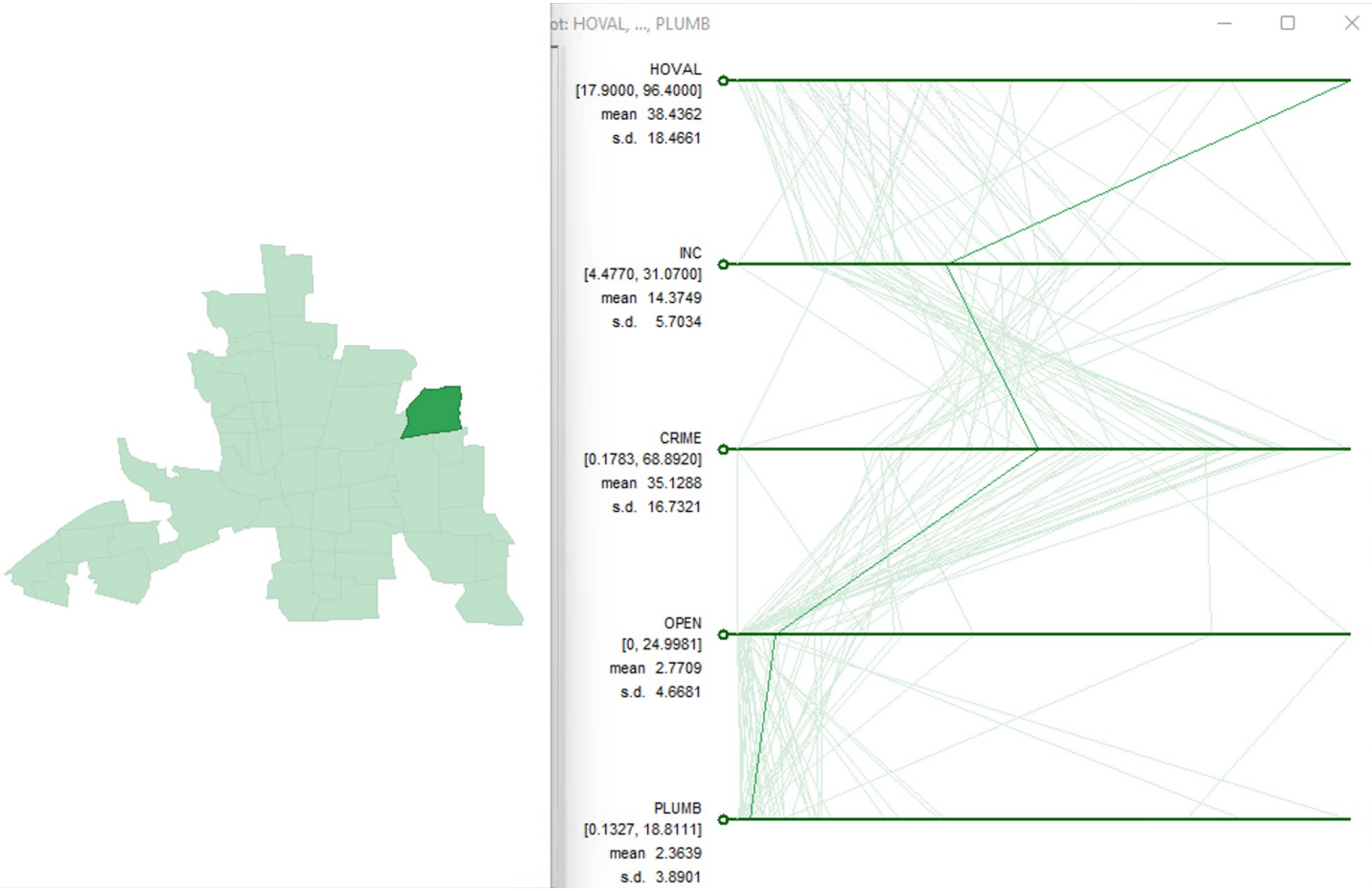
- The lower and upper lines of the "box" in the center of the plot window are the 25th and 75th percentiles of the sample. The distance between the top and bottom of the box is the inter-quartile range (IQR)
- The line in the middle of the box is the sample median. If the median is not centered in the box it is an indication of skewness
- The *whiskers* are lines extending above and below the box. They show the extent of the rest of the sample (unless there are outliers). Assuming no outliers, the maximum of the sample is the top of the upper whisker. The minimum of the sample is the bottom of the lower whisker.
- A symbol, e.g. a small circle, at the top and/or bottom of the plot is an indication of an outlier in the data. This point may be the result of a data entry error, a poor measurement or perhaps a highly significant observation
- The notches in the box are a graphic confidence interval about the median of a sample. A side-by-side comparison of two notched box plots is sometimes described as the graphical equivalent of a *t*-test. Box plots do not have notches by default

Other Visualization

- Parallel Coordinate Plots
- Conditional Plots

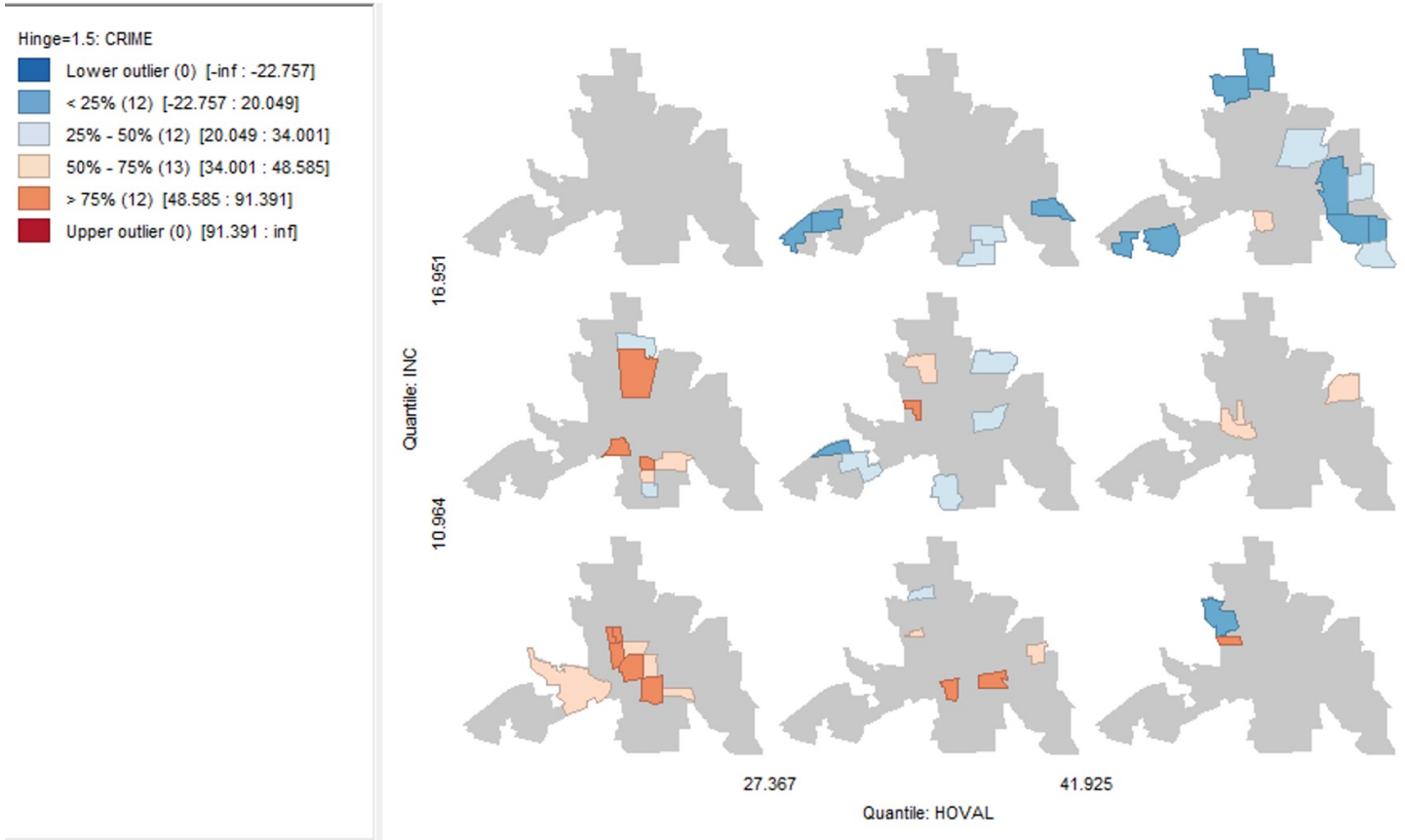
Parallel Coordinate Plots

- Multiple variables each with min-max scale
- Lines can be themed on colors



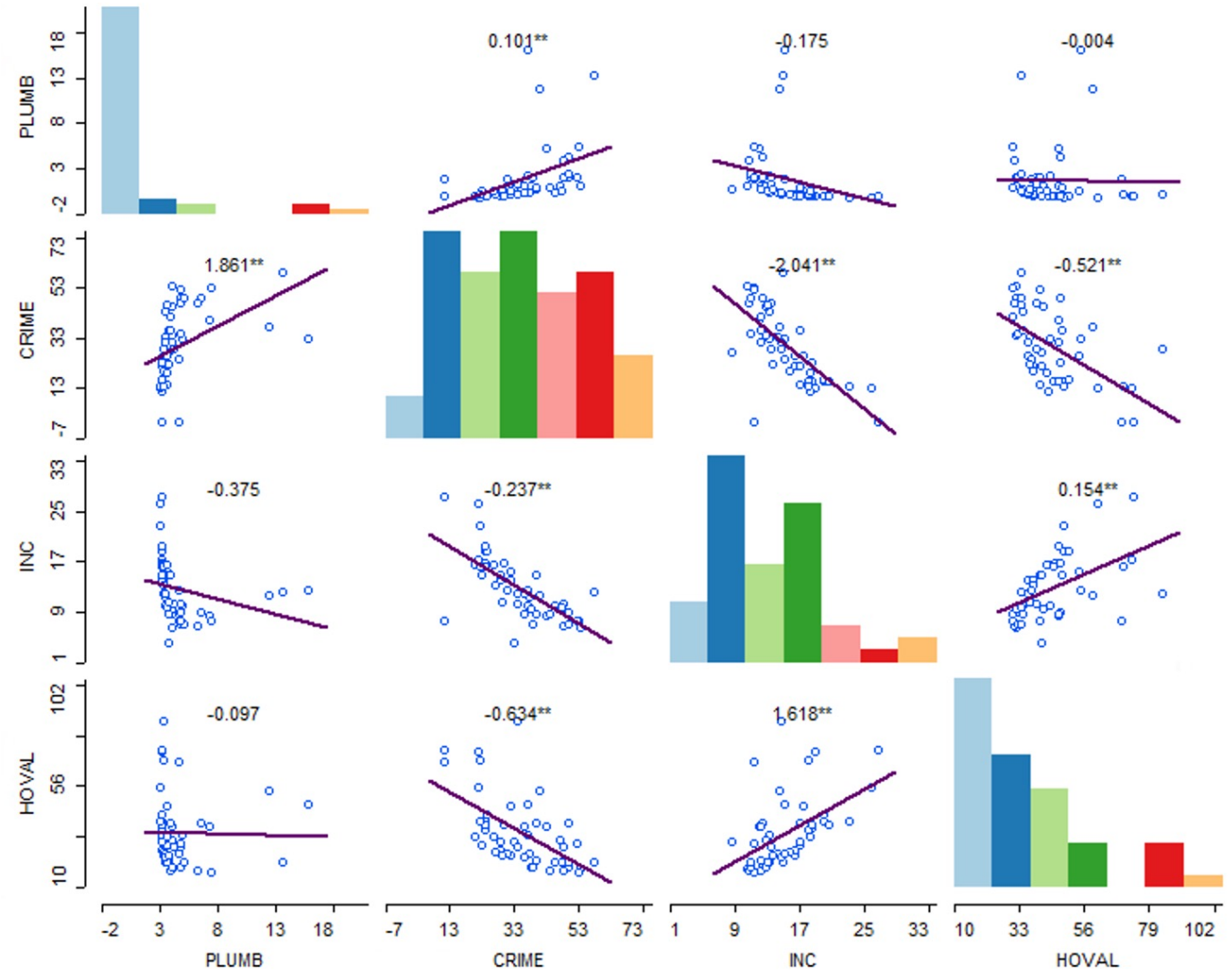
Conditional Plots

- Detect unexpected
- Check assumptions

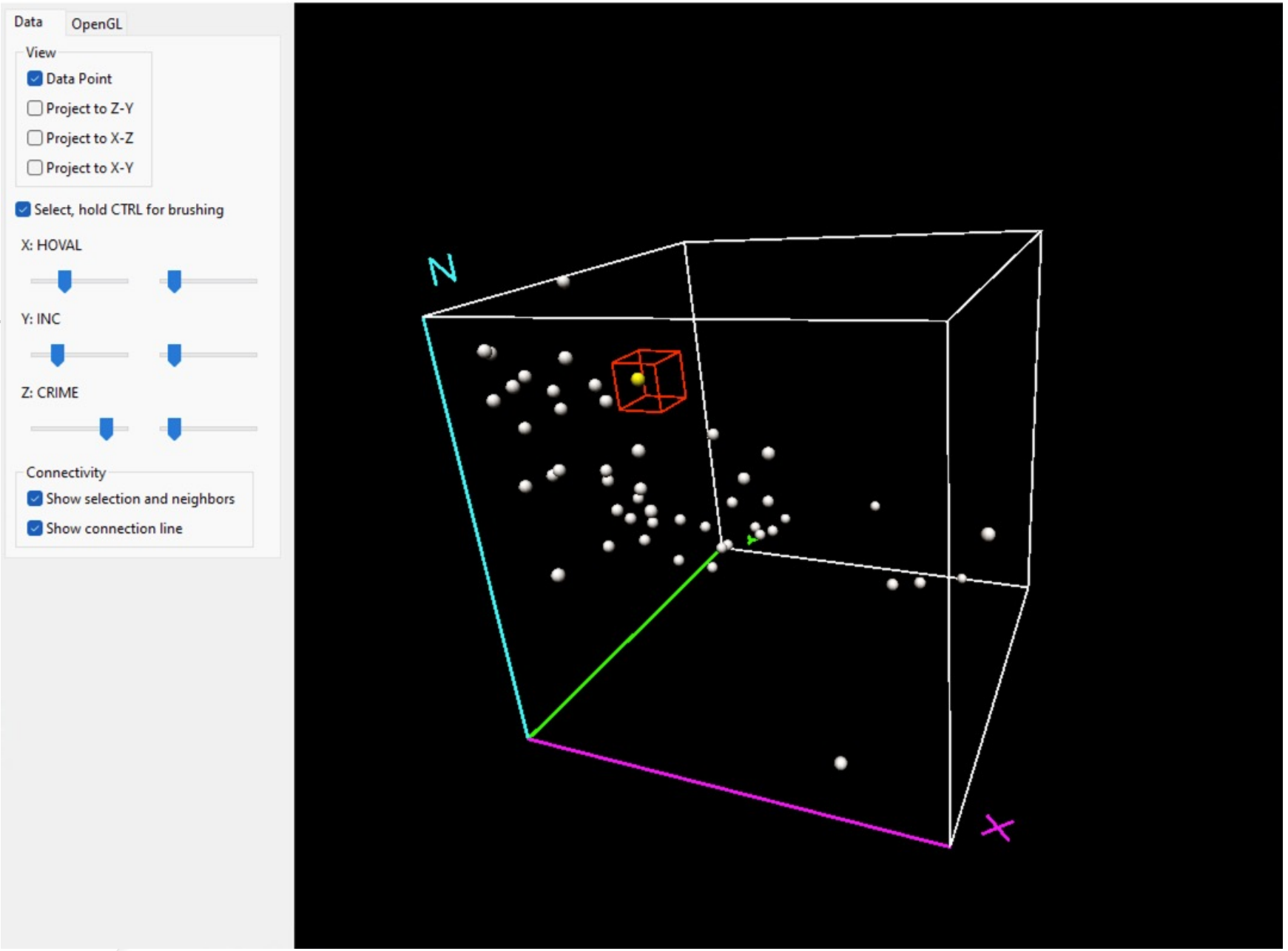


Scatter Plot Matrix

- Multiple Variables



3D Plots


Data
OpenGL
View
Data Point
Project to Z-Y
Project to X-Z
Project to X-Y
Select, hold CTRL for brushing
X: HOVAL
Y: INC
Z: CRIME
Connectivity
Show selection and neighbors
Show connection line
N
X

Dimensionality Reduction

- Data dimension is the number of variables for a measured theme/dataset
- Data with a high dimensionality is difficult to visualize
- Reducing the dimensionality of the data helps understand the intrinsic aspects of the dataset
  - Find structure within features
  - Aid in visualization
- The methods maximize information while minimizing differences between the original data and the new lower dimensional representation
- Principal Component Analysis is one such method

Principal Component Analysis

- Widely used method for dimensionality reduction
- The transformation between original data and the new lower dimensional representation is a linear projection
  - find a linear combination of the original features principal components.
  - The principal components will maintain as much as is possible the same variance as the original data
  - The principal components are uncorrelated (orthogonal)

Principal Component Analysis

- PCA major steps
  - Standardize the variables
    - Centre (deviation from the mean)
    - Scale (divide the deviation by the standard deviation)
  - Calculate the covariance matrix
    - Covariance – how 2 variables vary with each other
    - If you have more than 2 variables, then you gave more than one covariance (given variables x,y,z –> cov(x,y), cov(x,z), cov(y,z))
  - Calculate the eigenvectors and eigenvalues of the covariance matrix

$$Av = \lambda v$$

$$A \ - matrix$$
$$v \ - eigenvector \ for \ \lambda$$
$$\lambda - eigenvalue \ for \ v$$

Principal Component Analysis

- Recall matrix multiplication

$$\begin{bmatrix} A & B & C \\ D & E & F \\ G & H & I \end{bmatrix} * \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} AX + BY + CZ \\ DX + EY + FZ \\ GX + HY + IZ \end{bmatrix}$$

$$k \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} kX \\ kY \\ kZ \end{bmatrix}$$

$$\begin{bmatrix} A & B & C \\ D & E & F \\ G & H & I \end{bmatrix} * \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = k \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

A v λ v

Example and references

- GeoDa
  - http://geodacenter.github.io

- PCA overview:
  - Urška Demšar, Paul Harris, Chris Brunsdon, A. Stewart Fotheringham & Sean McLoone (2012): Principal Component Analysis on Spatial Data: An Overview, Annals of the Association of American Geographers, DOI:10.1080/00045608.2012.689236