GEOG 413/613

LECTURE 4

1

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
    - enables understanding and communication

2

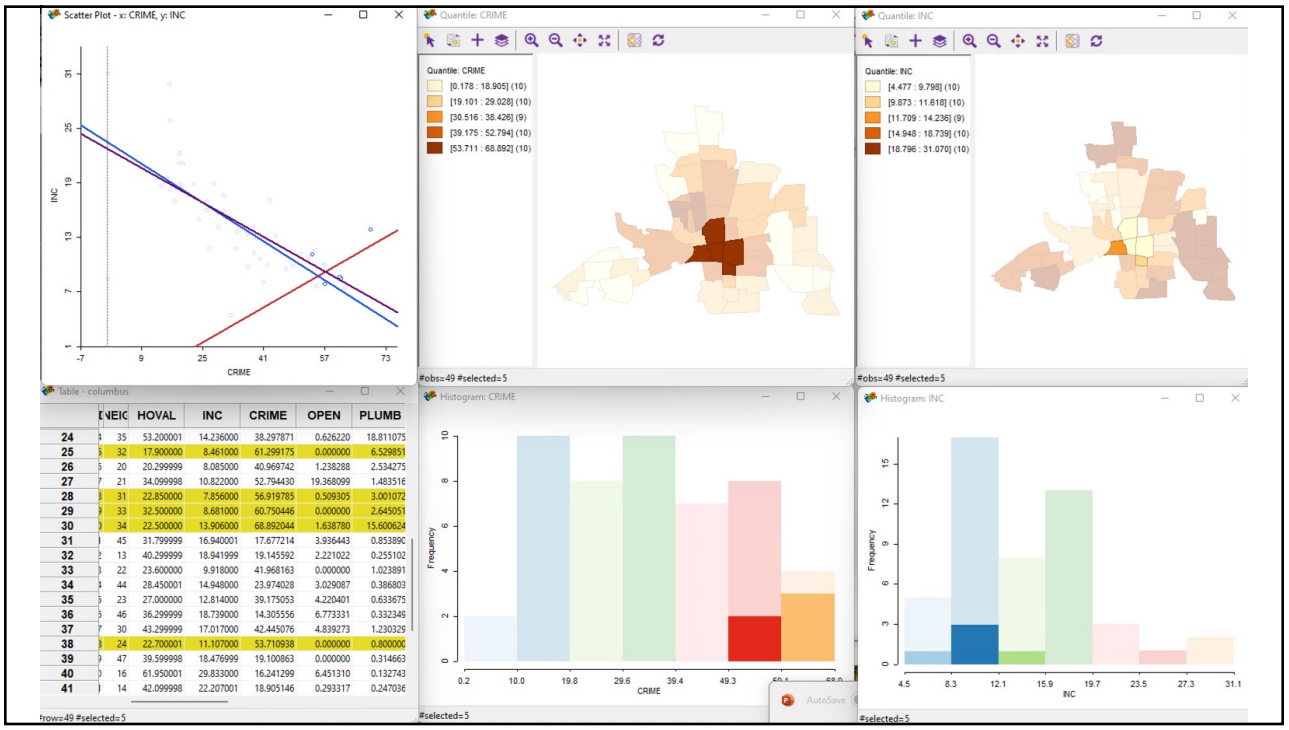
2

Exploratory Data Analysis

- Through EDA:
  - Explore data properties
    - distributions, data spread, central tendencies, etc
  - Dataset structure
    - How variables interactions with each other
- Visual and numeric tools are used

3

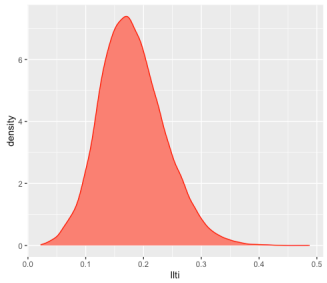
3

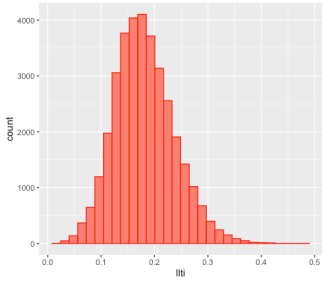


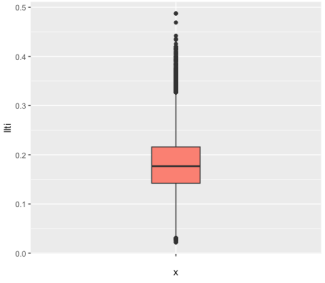
4

Single Continuous Variable EDA

- Common Tools
  - Can be visualized with a density plot
  - A histogram
  - Box plot



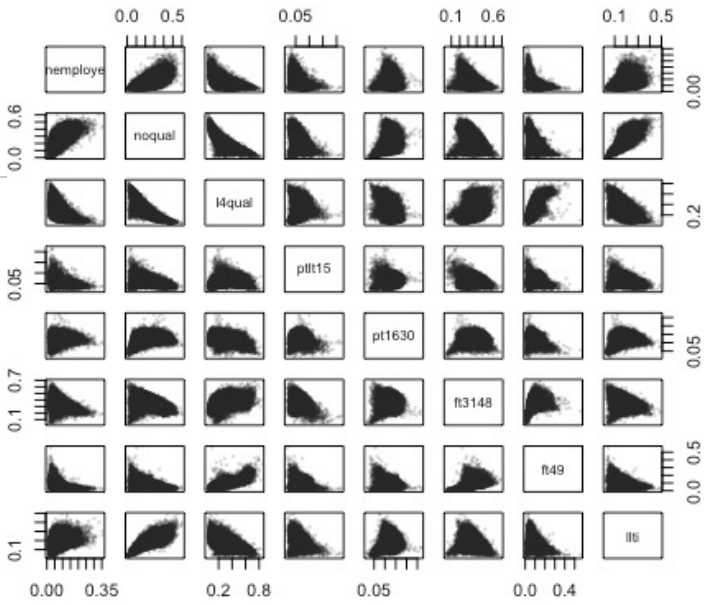




5

Multiple Continuous Variable EDA

- Common Tools
  - Visualization of correlations

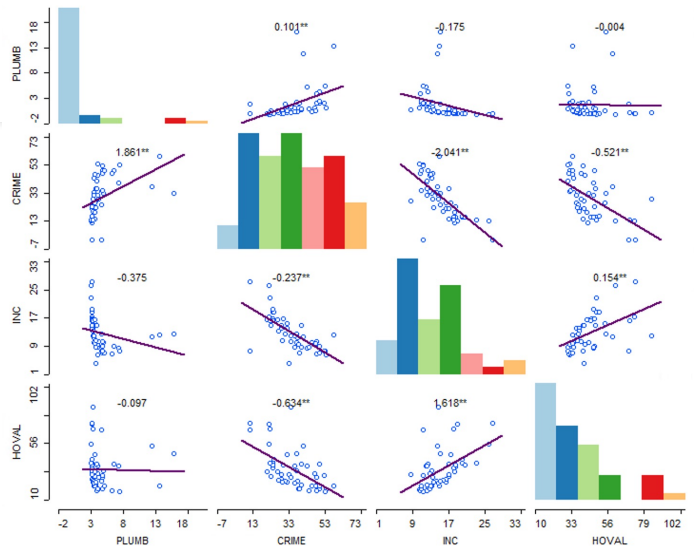


6

6

Scatter Plot Matrix

- Multiple Variables



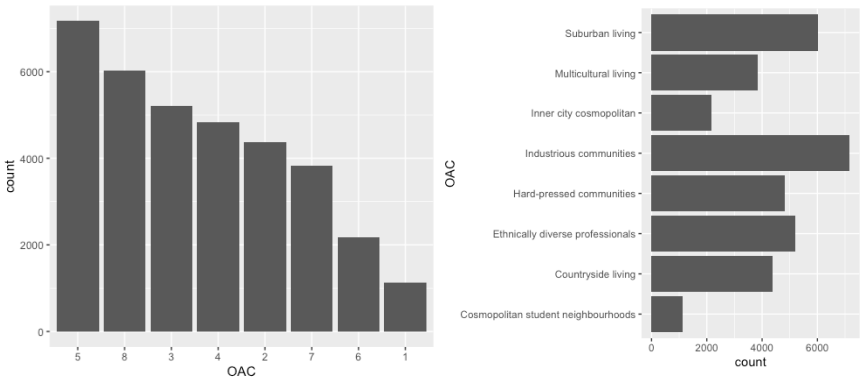
7

7

Single Categorical Variable EDA

Examined with Frequencies

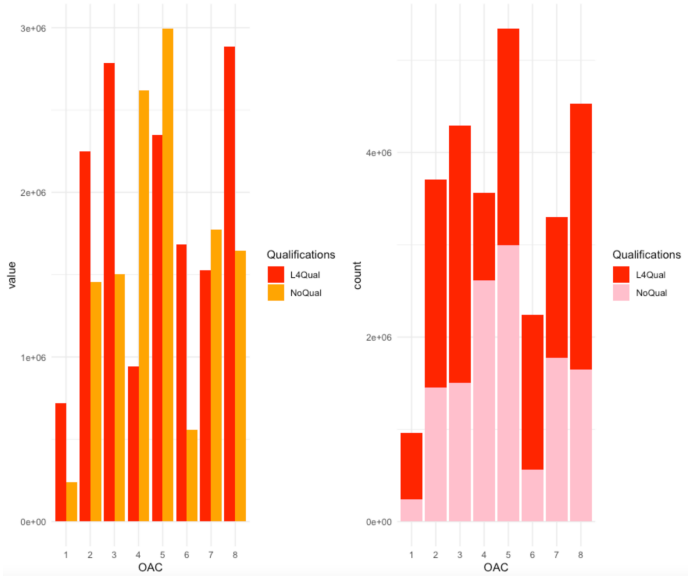
- Tables
- Bar graph



8

8

Single Categorical Variable EDA



9

9

Multiple Categorical EDA

Examined with tables and heat maps

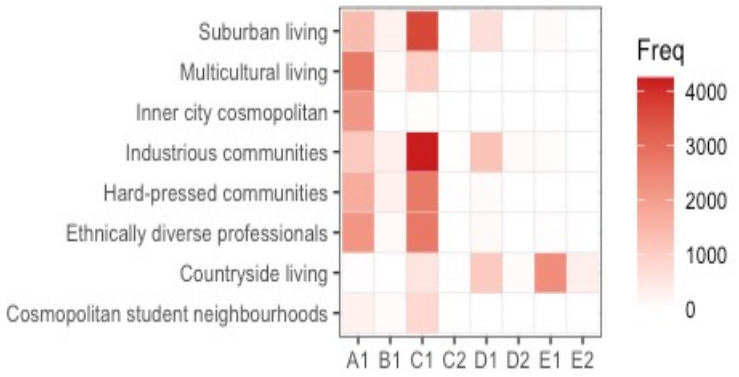
	rucll_code							
oac	A1	B1	C1	C2	D1	D2	E1	E2
Cosmopolitan student neighbourhoods	271	73	784	6	1	1	2	0
Countryside living	15	8	536	15	1082	80	2319	326
Ethnically diverse professionals	2158	123	2760	1	129	3	32	0
Hard-pressed communities	1706	281	2727	19	88	7	0	0
Industrious communities	1142	329	4252	50	1228	105	61	2
Inner city cosmopolitan	2121	3	46	0	0	0	0	0
Multicultural living	2724	131	983	0	0	0	0	0
Suburban living	1386	260	3636	3	661	1	76	0

10

10

Multiple Categorical EDA

Examined with tables and heat maps

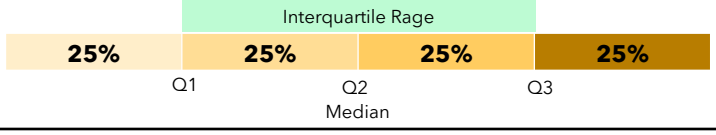


11

11

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

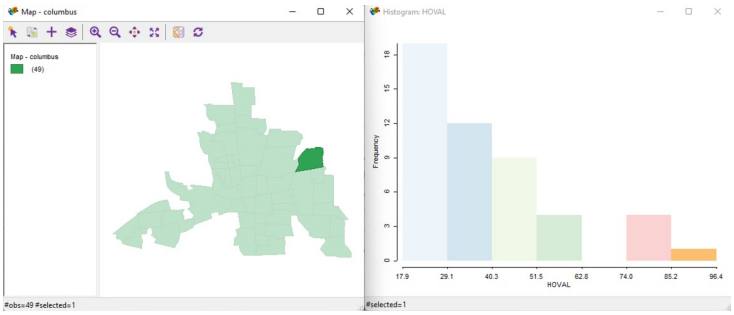


12

12

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

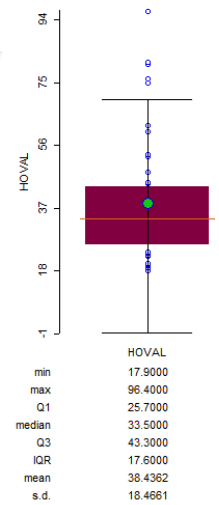


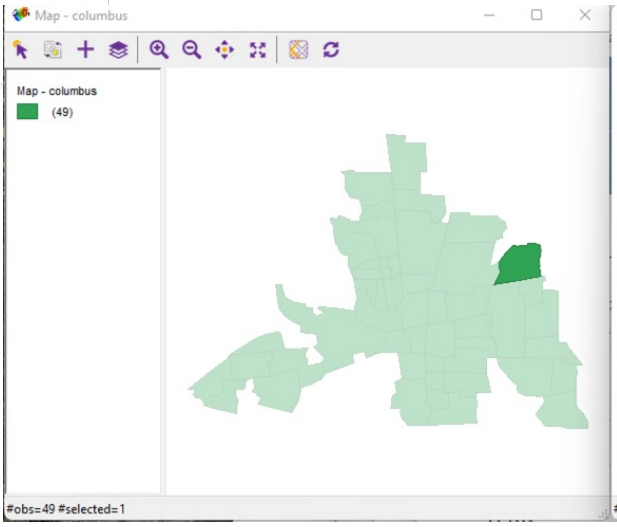
13

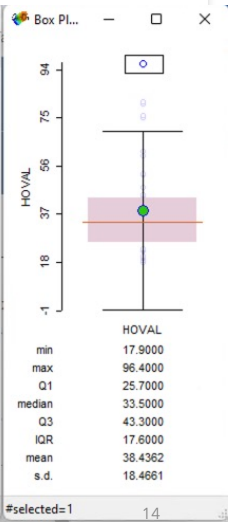
13

Outlier Detection

- Box plots







14

14

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

15

15

Other Visualization

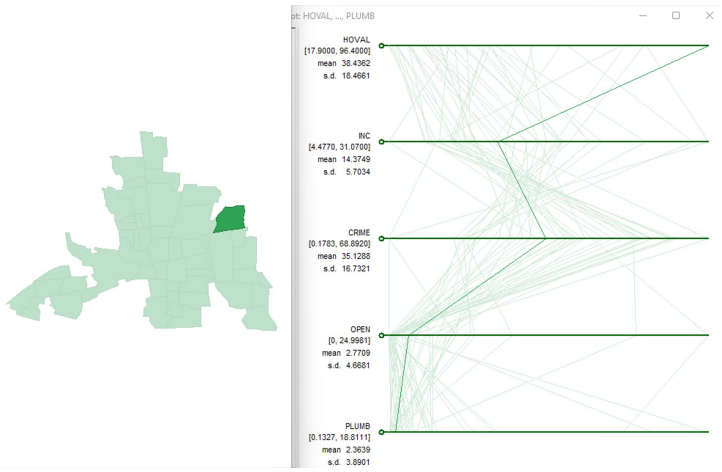
- Parallel Coordinate Plots
- Conditional Plots

16

16

Parallel Coordinate Plots

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

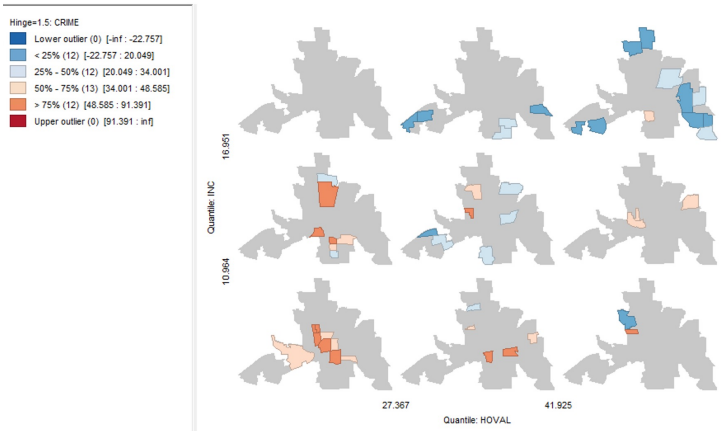


17

17

Conditional Plots

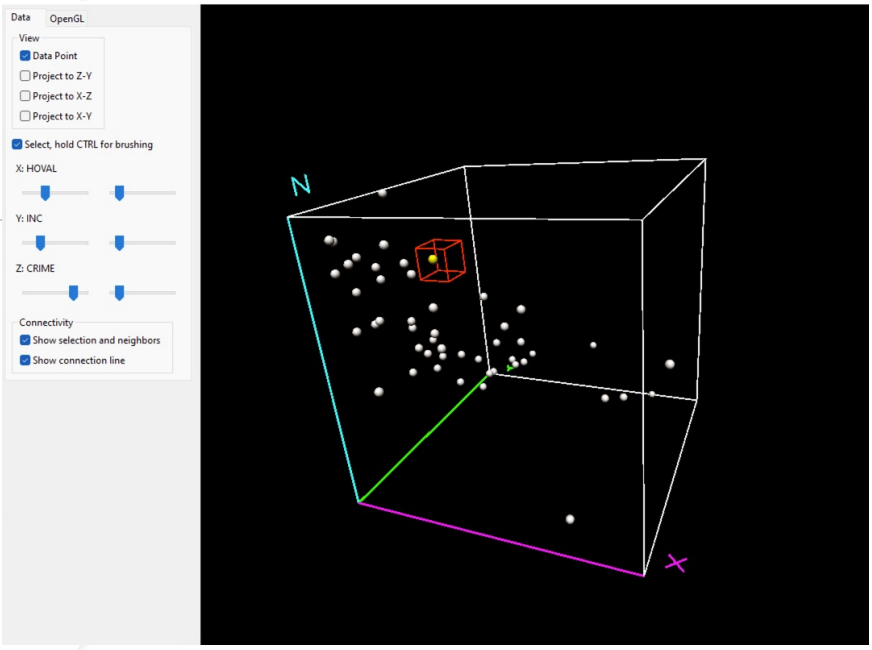
- Detect unexpected
- Check assumptions



18

18

3D Plots



19

19

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

20

20

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)

21

21

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

$$Av = \lambda v$$

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

22

22

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

23

23

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

24

24