Textbook Notes

Desription: This is the summary note from course textbook "Graphical Data Analysis", not from the course EDAV itself. Many contents are overlapped between these two but there are still a number of significant difference. This textbook note only covers the contents from Chapter 3 to Chapter 11 since Chapter 1 and 2 are introductions and contents after Chapter 12 are more likely to be summaries and introduction to R, which will not be covered in this file.

Textbook Notes
Chapter 3: Continuous Variable
- What to notice
- Plots
- Plot Options
- Model and Test

Chapter 4: Categorical Data
- Categorical Type
- What to notice
- Plot
- Model and Test

Chapter 5: Dependency
- What to notice
- Plot
- Plot options
- Model and Test

Chapter 6: Multivariate Continuous Data (Parallel Coordinate Plot)
- What to notice
- Plot Option
  - Format

Chapter 7: Multivariate Categorical Data (Mosaic Plot)
- What to Notice
- Plot Option
  - Plot Form
  - Others
- Model and Test

Chapter 8: Data Overview
- Just view
- Individual Display
- Multivariate Continuous
- Multivariate Categorical
- Graphics by Group
- Model and Test

Chapter 9: Data Quality
- Missing Data

- Outlier
  - Dealing with outliers
  - Possible Strategy
- **Chapter 10: Comparison**
  - Type of Comparison
  - Visualization
  - Principle
  - Model and Test
- **Chapter 11: Time Series**
  - Single Time Series
  - Multiple Time Series
  - Watch out
  - Alternative Plots
  - Model and Test

Chapter 3: Continuous Variable

What to notice

Type	Explanation or Comment
Assymetry(skew)	/
Outlier	/
Multimodality	Sometimes more than one peak
Gap	No data
Heap	Some values happen more frequently
Rounding	Only certain values (integer or some good-looking number) are found
Impossibilities	Zero possibility
Error	Low possibility

Plots

Type	Features and Comment
Histogram	See distribution The thinner the bin, the more gaps and heaps there are not for small data
Boxplot	Compare distribution, see outliers Good when there are outliers (compared with histograms) bad when multmodality exists
Dotplot	Look at the gaps
Rugplot	plot each individual as a line
Density Estimation	added to a plot compare distribution
QQ	compare data distribution to another distribution(usually normal distribution)

Plot Options

Where there is a skew, try to apply some transformations (like Box-Cox)

Name	Option
Binwidth	Integer is better needs good anchorpoints unequal binwidth not accepted
Bandwidth	For different density estimation
Scale	when by group, scale should be same

Model and Test

Type	Stat or Model
Mean	t-test
Median	Zheng, T. and Gastwirth, J. (2010). On bootstrap tests of symmetry about an unknown median. *Journal of Data Science*, 8:397–412.
Symmetry	bootstrap
Normality	nortest
Density Estimation	logspline and other R packages
Multimodality	diptest and other R packages

Chapter 4: Categorical Data

Categorical Type

Type	Comment
Single Category	rarely used but exist
Nominal (no order)	/
Ordinal	order must be preserved
Discrete	order must be preserved

What to notice

Type	Comment
Unexpected Pattern or Results	/
Uneven Distribution	Sometimes results are on specific results
Extra Categories	'M', 'F' maybe others
Unbalanced Experiments	/
Large Number of Categories	/
Refusal, Error, Missing ...	/

Plot

Barplot

Pie (not recommended)

Dynamite plot

Model and Test

Type	Stat or Comment
Test by simulation	χ^2 test
Eveness of distribution	χ^2 test
Fit discrete distribution	χ^2 test

Chapter 5: Dependency

What to notice

Type	Comment
Casual Relationship	Linear / non-linear
Associations	no casual but just association
Outlier	Outlier / group of outliers
Cluster	group of cases
Gap	some particular combinations do not occur
Barrier	some combination area is not possible
Conditional Relationship	relationship changes for different condition

Plot

Main	Scatterplot
Levels	contour, hdrcde
Line	geom_smooth, stat_smooth
Comparing Groups	facet_wrap
Pair of values	ggpairs, spm, splom

Plot options

Type	Comment
Points	Small points: hardly seen, easy to group Large points: easy to see, overlap
Point symbol	available for small data
Alpha blending	overlap goes down, outlier detect goes down interactive is better
Color points	only use when it can show clusters
Splom	Versatile

Model and Test

Type	Stat or Model
correlation	linear regression
regression	linear regression + confidence interval
smoothing	loess
bivariate density estimation	kde, kde2d, bkde2d
outlier	NO WAY

Chapter 6: Multivariate Continuous Data (Parallel Coordinate Plot)

What to notice

Type	Comment
Gap/Concentration	/
Skew	/
Outlier	Outlier / group of outliers
Clustering	Visualize than just accept

Plot Option

Type	Option
Alignment	max, min, median, mean
Scaling	uniminmax, IQR
Outlier	Remove Outlier, Trim, Restrict Plots, logarithm
Variable Order	sort by variance, mean, IQR

Format

Type	Option
Display type	showpoints, boxplot
Missing	include, exclude (default)
Aspect ratio	/
Orientation	Horizontally / Vertically to avoid overlap
Lines	Amount ++, thinner
Color	by group
Alpha blending	lessen overplotting problem

Chapter 7: Multivariate Categorical Data (Mosaic Plot)

Plotting this is **DIFFICULT**

What to Notice

Most frequent subgroup

Compare between subgroup

Pattern of subgroup

Look at the residual after modeling

Plot Option

Plot Form

Situation	Recommendation
ordinal data	classic mosaic
Dependence	Multi-bar
many combinations	fluctuation diagram
compare rates	doubledecker, same binsize plots
missing combination	same binsize plots
Compare distribution	rmb plot

Others

Option Type	Comment
Ordering	binary dependent variable should be the last ordinal must be in order
Display	Big, less color, no label, captions and annotations are important
Aspect Ratio	For diagram and binsize plots, they should be square For others, they should be tall and thin
Gaps	by hierarchy
Color	by subgroup or residual

Model and Test

Case	Stat and Model
Association	χ^2 Test
Small number of variables	logistic regression
Binary independent	logistic regression

Chapter 8: Data Overview

Just view

Function: summary, describe, whatis

Function: str() to see the type of data

Individual Display

barplot, histogram to get distribution and feature

Multivariate Continuous

scatterplot matrix or coordinate plots to understand principles between features

heatmap and glyph are also options

Multivariate Categorical

Mosaicplot

Multiple bar charts

Graphics by Group

Trellis: several kinds of group

Group plots: one group

Model and Test

Type	Stat and Model
Transformation	Box-Cox
Association	χ^2 Test
Discrimination	SVM

Chapter 9: Data Quality

Missing Data

Visualization: mi package for taking a look at missing data. Or extract::visna

MAR or MCAR: compare data missing subsets, using fluctile to see the missing pattern between groups

Missing Variable handing: some data are tricky, using "99" for missing

Outlier

Type	Handling
Univariate	Boxplot outlier default: 1.5 * IQR prior knowledge needed skew the transformation
Multivariate	Scatterplot, parcoord, split
Categorical Outlier	Fluctuation Diagram

Dealing with outliers

Obviously wrong: Discard or Correct

Little effect on performance: keep

Practical modeling: weighted linear combination

Possible Strategy

2-dim distribution -> potential outlier examination -> high-dim outlier -> outlier in the subset

Chapter 10: Comparison

Type of Comparison

Type: specific, general, different levels

Comparing: population, variable, source, group, condition, measurement, standardization

Visualization

situation	method
compare to a standard value	histogram + vertical line + confidence interval
new data vs old data	2 histogram
subgroup comparison	Boxplot / density estimates / confidence interval plot
Time series comparison	line + color between difference
subsets	trellis

Principle

Graph size should be same

Common Scaling

Alignment

Color is better than shape

Model and Test

Situation	Stats and Model
Mean	t-test
Complex comparison	linear models
Rate	Proportional odd model
Non-parameteric	Wilcoxon for mean, Kruskal-Wallis for variance

Chapter 11: Time Series

Single Time Series

option	Consider
Symbol	Point or line or bar
Scale	Min, max, zero included?
Aspect Ratio	Trend is 45 degree
Gaps	fill gap or not

Multiple Time Series

Multiple Time Series

Situation	Choice
same population	draw each one independently all in one plot
subgroup	If scale varies a lot, use multiple graphs transformation is another choice
many series	Trellis

Watch out

Title	Content
Data definition	watch out the change of definition as time goes
Length of Time Series	short term makes long trend obsecure different term should be different scale
Regular vs Irregular	use special packages to let irregular time series be same time gaps
Outlier	not real outlier but just outlier in a part scale adjustment interactive zoom in or out
Forecasting	shaded region + gap + dot line
Patterns	easy to overlook features inconsistent with supposed pattern

Alternative Plots

bar plot

parcoord

calendar plot

Model and Test

Type	Stat and Model
single time series	ARIMA, GARCH, decomposition
short irregular time series	Smooth
Multivariate time series	NO WAY