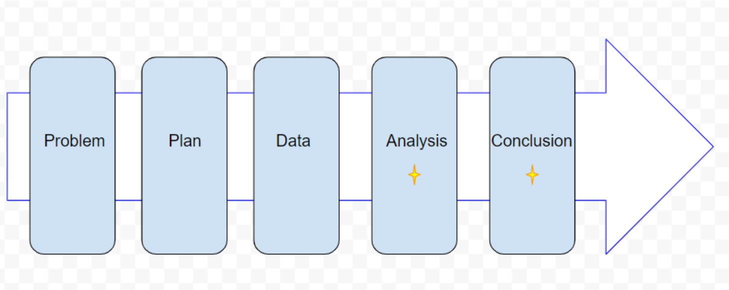
STAT 938 Consulting Workshop: ggplot2

Table of Contents

1. Grammar & graphics
2. Best Practices in ggplot2
3. Extensions of ggplot2
4. Case Studies

Background

- **PPDAC:** Organizational framework for statistical projects
- **Graphs** involved in Analysis and Conclusion



PPDAC Framework

Grammar and Graphics

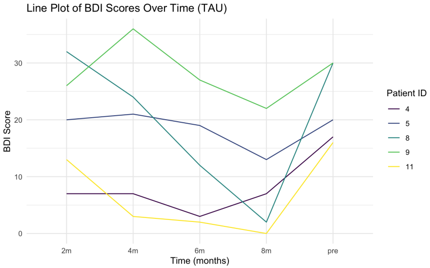
Grammar and Graphics

Grammar and Graphics

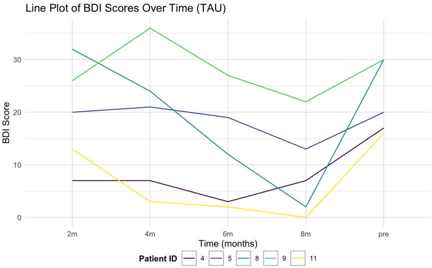
Themes in ggplot2

- **Themes** in ggplot2 control the non-data elements of your plot, such as the background, grid lines, text, and legends.
- **Default Themes**: ggplot2 comes with several built-in themes like `theme_minimal()`, `theme_classic()`, and `theme_bw()`.
- **Custom Theme**: Combining different theme elements to fit presentation or publication needs.

Theme Comparison and Custom Theme Function



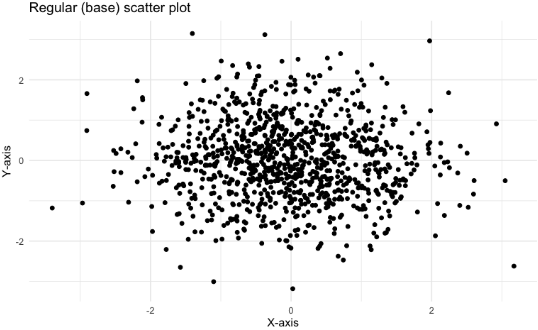
Default Theme



Custom Theme

Dealing with Overplotting

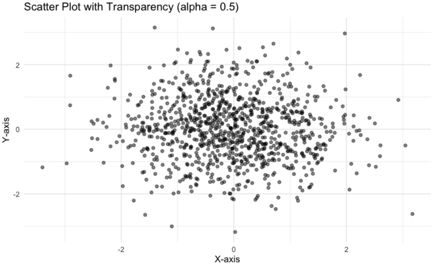
- Overplotting occurs in large datasets where points may overlap, preventing accurate assessment of the distribution of the data.
- **Example Application:** Simulated (x,y) point data from the standard normal distribution.



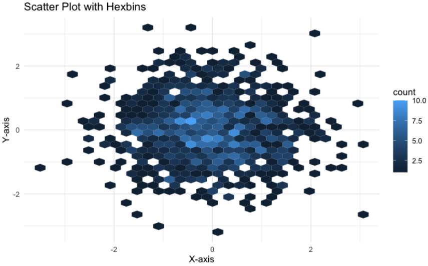
Regular Scatter Plot

Dealing with Overplotting

- **Transparency**: Reduce the opacity of points using the ‘alpha‘ parameter to make overlapping points visible and highlight areas of high density.
- **Hexbins**: Use hexagonal binning to aggregate points into hexagonal cells, providing a clear visualization of point density in large datasets.



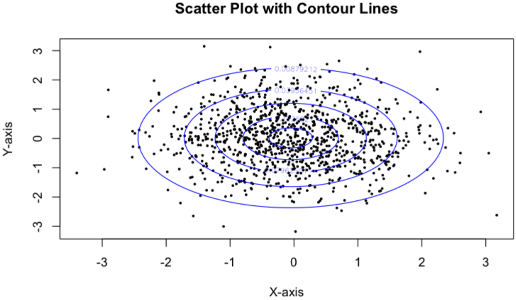
Transparency



Hexbinning

Dealing with Overplotting

- **Contour Lines**: Contour lines represent the density of points in a scatter plot, helping to visualize the distribution and concentration of data points.



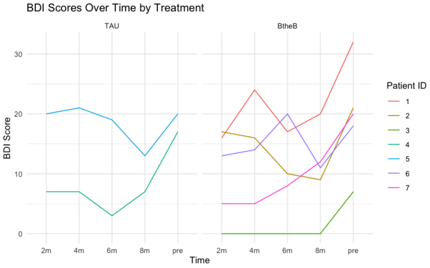
Contoured Scatter Plot

Facets

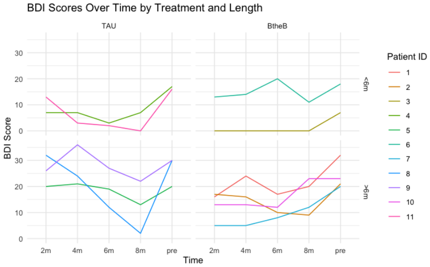
- **Facets** are used in ‘ggplot2‘ to create multiple subplots that each display a subset of the data.
- They allow for easy comparison of different subsets within a single visualization.
- Faceting is useful for exploring patterns across different levels of a categorical variable.
- **Example Use Case**: Comparing BDI scores across different treatments or time periods.

Facets

- Two main functions:
  - facet_wrap(): Creates a series of plots wrapped into a specified number of rows and columns.
  - facet_grid(): Creates a grid of plots based on two categorical variables.



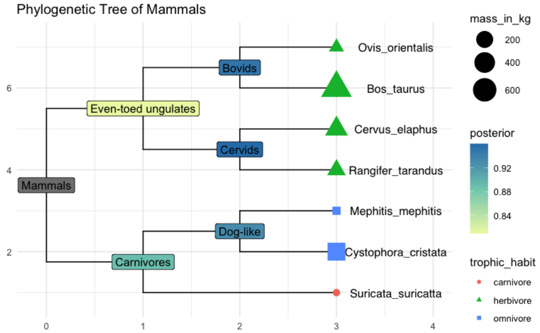
Facet Wrap



Facet Grid

ggtree

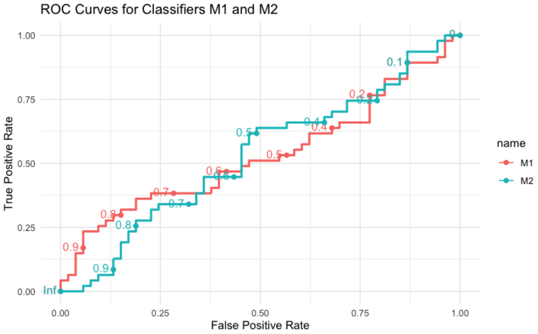
- **ggtree** is used to visualize and annotate phylogenetic trees.
- Phylogenetic or evolutionary trees are diagrams that represent the evolution relationships among various species based off their characteristics.



Phylogenetic Tree Example

plotROC

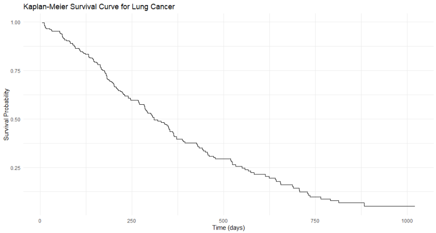
- **plotROC** is used for visualizing and comparing the performance of binary classifiers using Receiver Operating Characteristic (ROC) curves.
- Compares False Positive Rate (1 - specificity) against True Positive Rate (Sensitivity) at various thresholds.



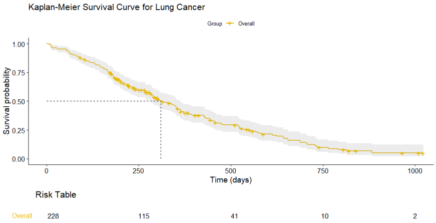
ROC Curve example

Kaplan Meier (KM) Curves

- **KM Curves** estimate the probability of survival over time for a group of subjects
- **R libraries:** survminer & survfit
- **Uses:** Epidemiological/Public health studies



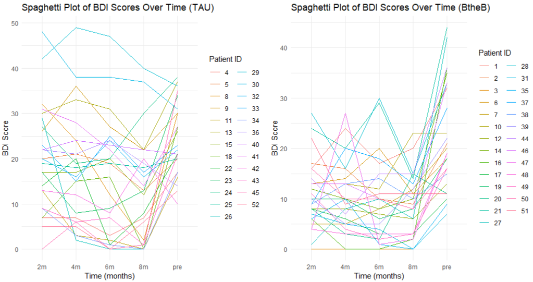
KM Curve with ggplot2



KM Curve with ggsurvplot from survminer

Spaghetti & Lattice Plot

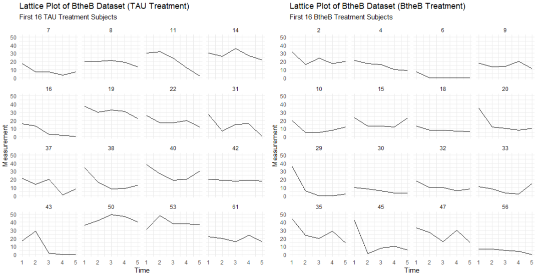
- **Spaghetti Plots:** individual trajectories or lines are plotted over time or some other continuous variable
- **Uses:** Longitudinal data/Clinical studies/Cohort/Case-control studies



ROC Curve example

Spaghetti & Lattice Plot

- **Lattice Plots:** multiple plots in a grid-like structure, each plot representing a subject
- **Uses:** Longitudinal data/Clinical studies/Cohort/Case-control studies



Lattice Plot of BtheB Dataset (TAU Treatment)

First 16 TAU Treatment Subjects

Lattice Plot of BtheB Dataset (BtheB Treatment)

First 16 BtheB Treatment Subjects

ROC Curve example

Conclusion: ggplot2

- **Grammar of Graphics:** Structured layers
- **Good practices**
- **Extensions** animations, ROC, etc.
- **Public Health applications** Epidemiological/Case studies, Survival analysis (KM Curves), Clinical trials, etc.
- **Code** available in GitHub repository