Chapter 1

Introduction to Multivariate Statistics

HUDM 6122

Introduction

Definition: Multivariate Statistical Methods *are a collection of statistical methods used for the analysis of data where several measurements are made on each observational unit.*

- *The measurements are called* variables.
- *The observational units are simply called* observations (or cases).

Most datasets we encounter nowadays are multivariate in nature. We may have tens of responses by subjects to a questionnaire or test. We might be interested in the performance of students across several disciplines, etc. In many situations like these, univariate statistical methods may not apply.

Types of Multivariate Methods

The *analysis of interdependence*. The multiple measurements taken on each observation are almost always dependent (correlated); if they were not, then analysis would be simple. We often want to gain a better understanding of this interdependence.

- Principal components analysis: try to find linear combinations of the variables that describe the majority of the variation in the sample. Essentially, we are trying to find a smaller set of measurements that still contain most of the information that is in the original data.
- Factor analysis: identify underlying latent constructs that describe the interdependence among the variables.
  - Exploratory factor analysis: how many constructs are measured by the variables? What variables measure which constructs, etc.
  - Confirmatory factor analysis: used to test prior theories about how the variables measure one or more latent constructs.
- Multidimensional Scaling: find a low-dimensional graphical representation of the observations and variables. Think about drawing a map of the variables and observations.
- Cluster analysis: find how observations or variables are grouped together. Group objects with other similar objects, where similarity is defined in terms of the variables.

Types of Multivariate Methods

Understanding dependence. In regression analysis we are interested in figuring out if one variable (the response variable) is dependent on a set of explanatory variables. Multivariate methods expand multiple regression methods to understand the dependence of several response variables on a set of explanatory variables.

- Multivariate regression: several *x*'s and several *y*'s.
- Structural equation models: several *x*'s and several *y*'s measuring a few latent constructs (e.g., IQ, depression, etc.)
- Canonical Correlation: several *x*'s and several *y*'s. Try to find a reduced number of "canonical" variables that describe all of the dependence.
- Multivariate analysis of variance (MANOVA): how does the distribution of measurements vary by group or treatment?
- Multivariate analysis of covariance (MANCOVA): how does the distribution of measurements vary by group after we control for several covariates (e.g., gender, race, SES, etc.)?
- Discriminant analysis and Classification: Describe group differences or predict group measurement in terms of the multiple measurements.
    - Fisher's Discriminant Analysis
    - Logistic regression.
    - Classification trees.
    - Support vector machines.

Example

A psychiatrist has developed a 40-item questionnaire that is given to patients entering the emergency room with psychiatric problems. Each question asks the patient how often they have suffered some specific symptom, e.g., how often do you feel overwhelmed ... The patient responds on a 3-point scale (0 = never; 1 = occasionally; 2 = often). The psychiatrist also records gender, race, psychiatric history, medication history, age of the patient, and whether or not they have attempted suicide in the past.

Below are examples of multivariate analyses that could be accomplished with this data:

- MANOVA: Are there differences in the responses to the 40 questions between suicide attempters and non-attempters?
- MANCOVA: After controlling for suicide attempt, are there gender differences in response behavior?
- Discriminant/Classification analysis: try to describe the difference between suicide attempters and non-attempters in terms of their responses to the 40 questions. If there are differences, use the 40 items to identify individuals that might be at risk.
- Factor analysis: how many different constructs are measured by the 40 items? For example, is there one factor that appears to be measuring depression and another measuring panic?
- Cluster analysis: are there distinct groups of questions? Are there distinct groups of patients?

Most of what we will focus on in this course concerns quantitative measurements. The example above has ordinal measurements, but the types of analyses follow through.

Course Objective

By the end of the semester you should be able to:

- Understand the various multivariate statistical methods and know which one to use for a given problem, the assumptions behind the multivariate methods, and the limitations of each method.
- Utilize statistical software to perform the appropriate multivariate statistical analysis and how to interpret the output of the software.

Note:

All these methods are based on the *multivariate* normal distribution, which needs to be introduced first. However, before that we need to review linear algebra (vectors, matrices, determinants, …).

Data Layout

- In Multivariate Data, there will be *n* experimental or sampling units, each having *p* variables/characteristics being observed/measured
- The data will be structured in an array of *n* rows (units) and *p* columns (variables), which is labelled as **X**
- x_{jk} represents the measurement of the k^{th} variable on the j^{th} unit

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,p-1} & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2,p-1} & x_{2p} \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n-1,1} & x_{n-1,2} & \cdots & x_{n-1,p-1} & x_{n-1,p} \\ x_{n1} & x_{n2} & \cdots & x_{n,p-1} & x_{np} \end{bmatrix}$$

Descriptive Statistics

- Means:

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}\,, \qquad k = 1, \ldots, p$$

- Sums of Squares:

$$w_{ik} = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)\,, \qquad i = 1, \ldots, p;\ k = 1, \ldots, p$$

- Variances:

$$s_k^2 = s_{kk} = \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 = \frac{w_{kk}}{n}, \qquad k = 1, \ldots, p$$

(Note dividing by n, not n-1 for now)

- Covariances:

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) = \frac{w_{ik}}{n}, \qquad i = 1, \ldots, p;$$

$$k = 1, \ldots, p$$

(Note dividing by n, not n-1 for now)

Descriptive Statistics

- Correlations:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}s_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2 \sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2}} = \frac{w_{ik}}{\sqrt{w_{ii}w_{kk}}}, \qquad i = 1, \dots, p; k = 1, \dots, p$$

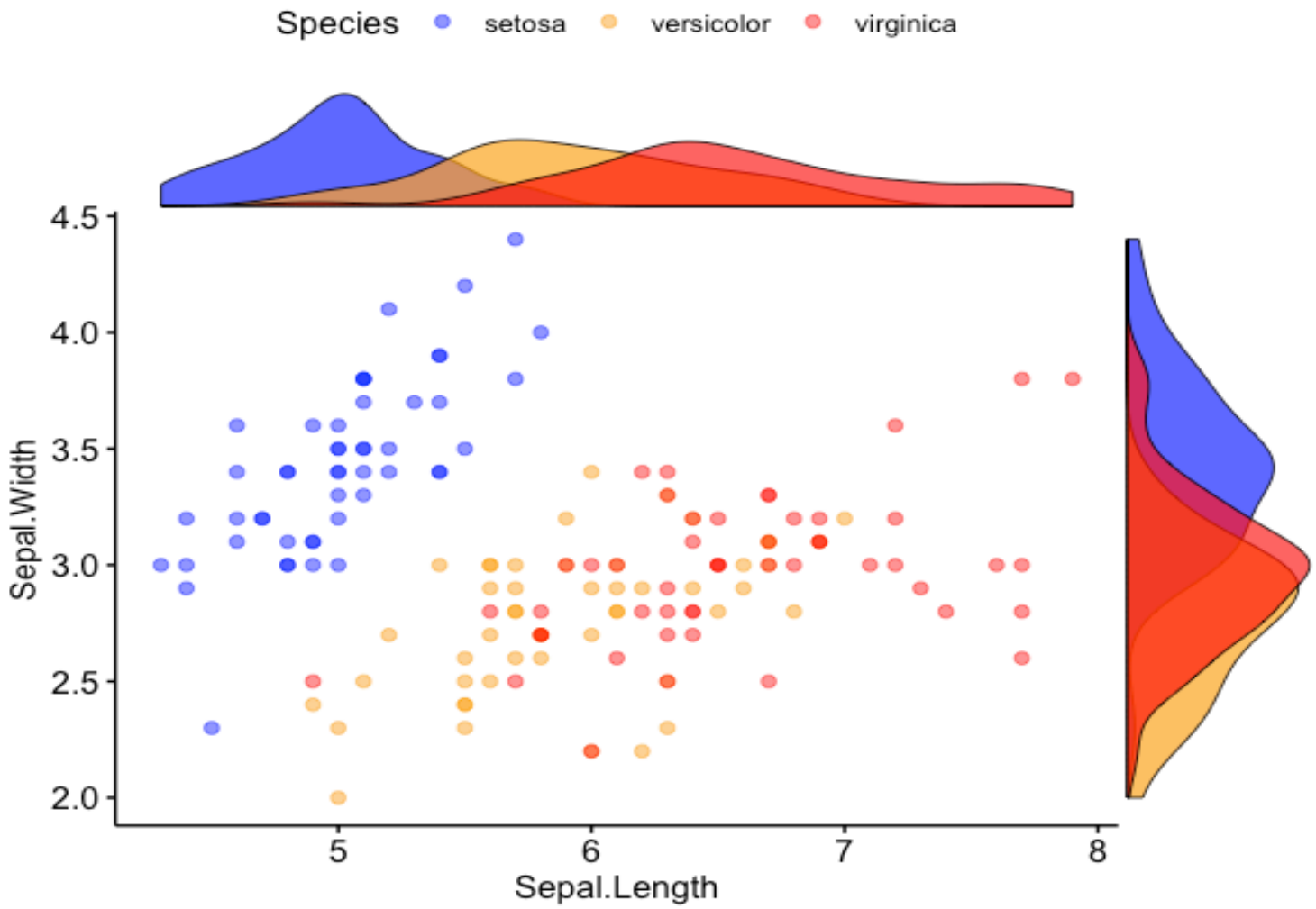
- Mean (column) *p*x1 vector:

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

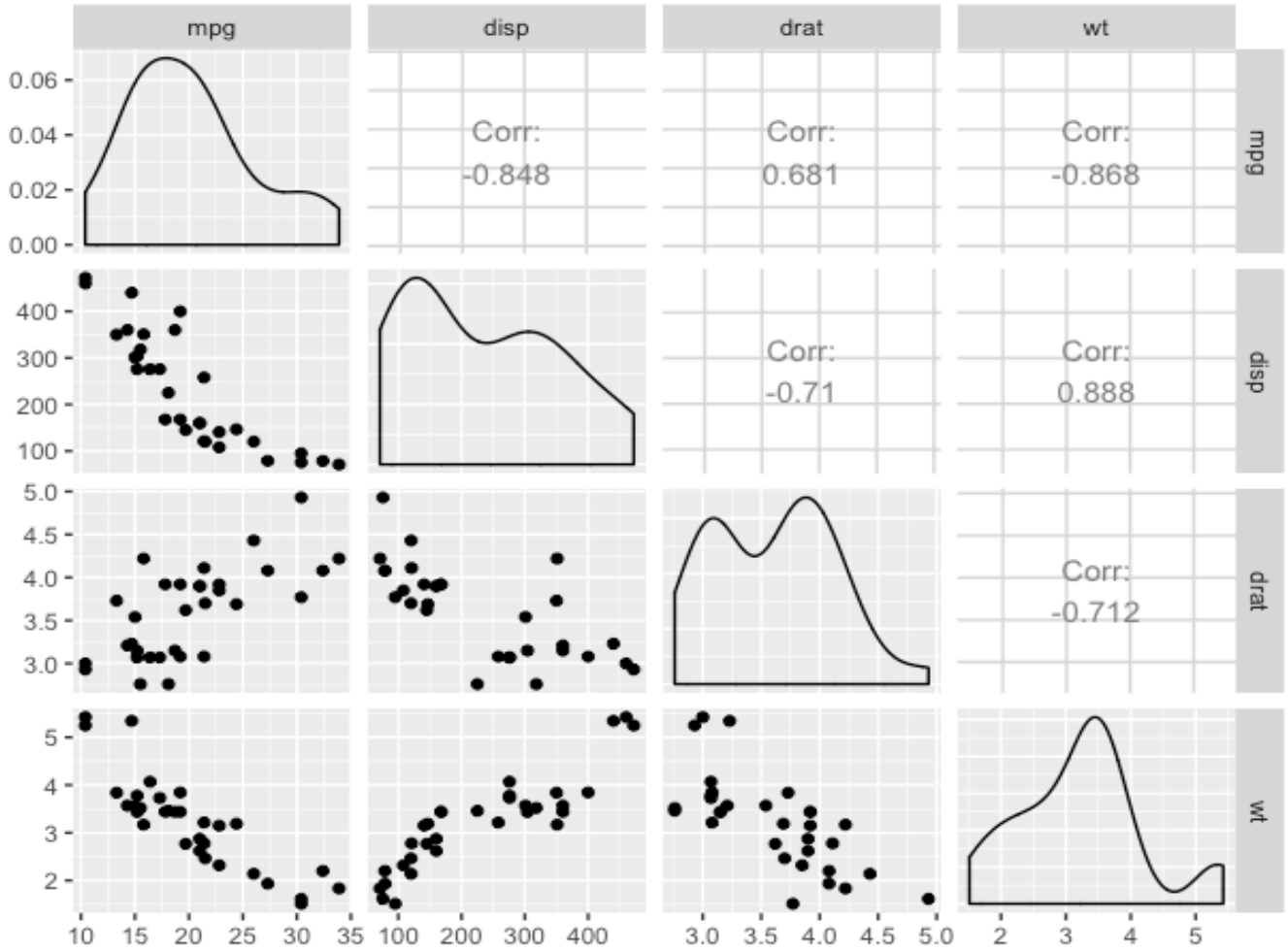
- Variance-Covariance *p*x*p* matrix:

$$S_n = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix}$$

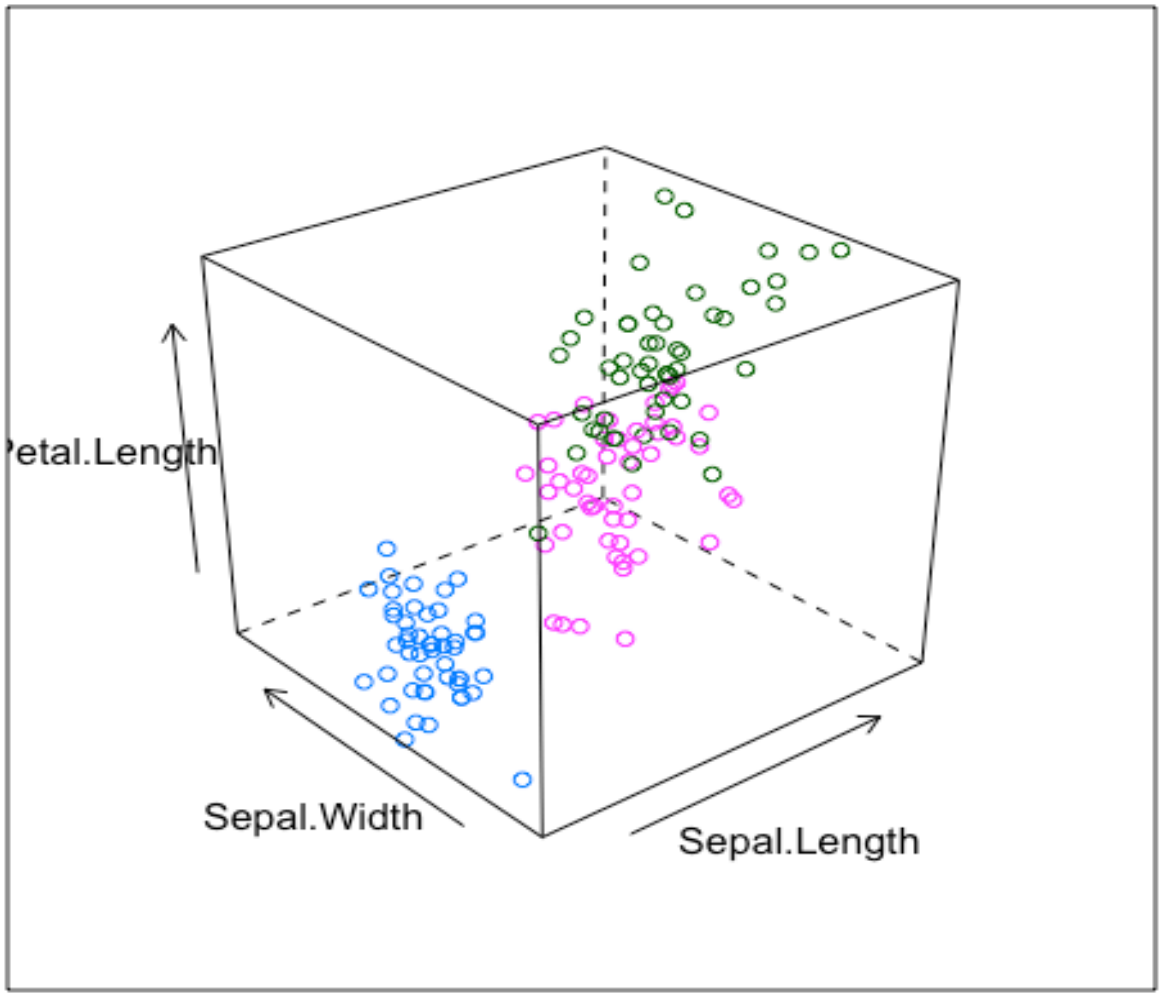
Scatterplot with Marginal Density plot – Iris dataset



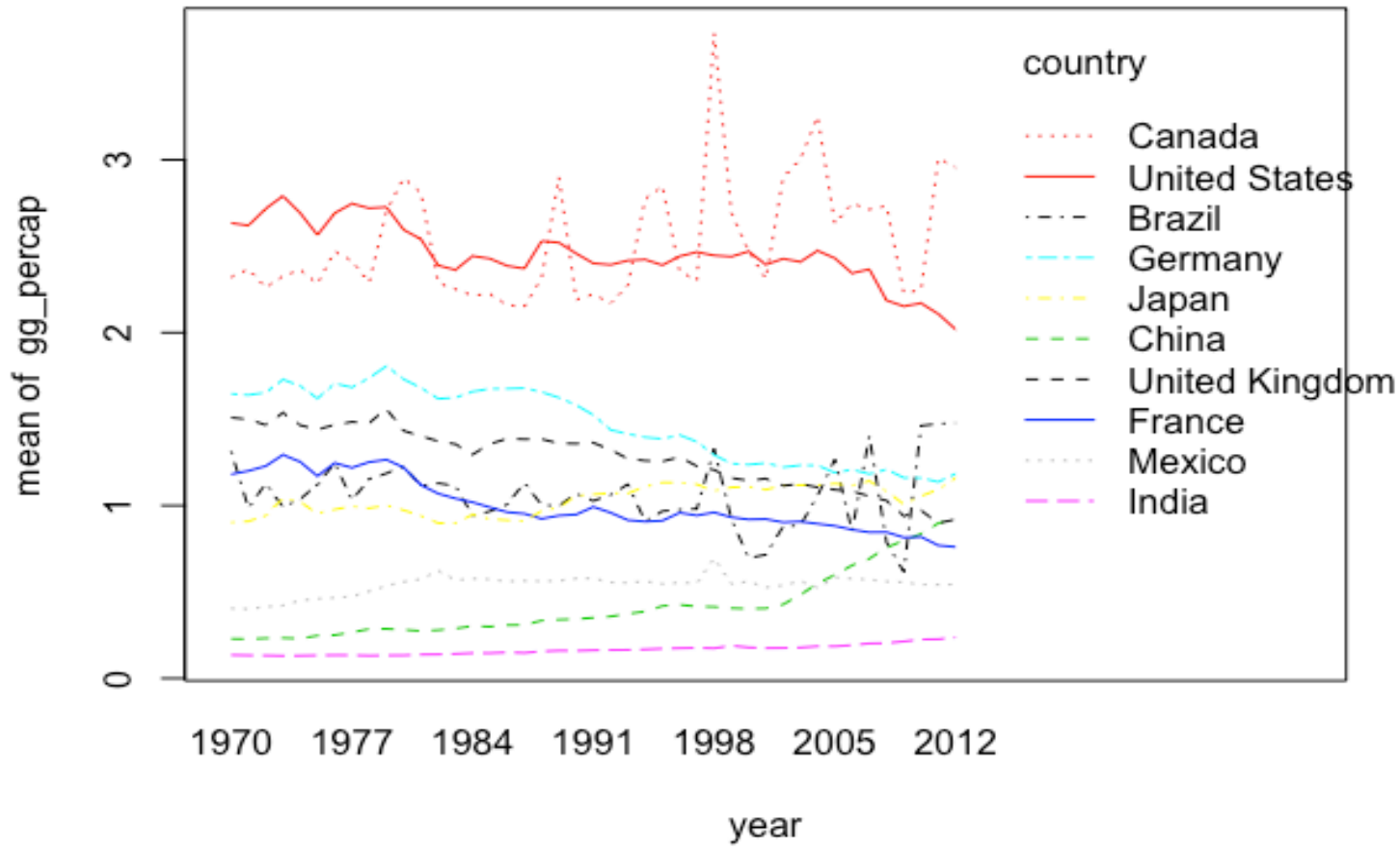
Scatterplot Matrix with Densities – MT cars data



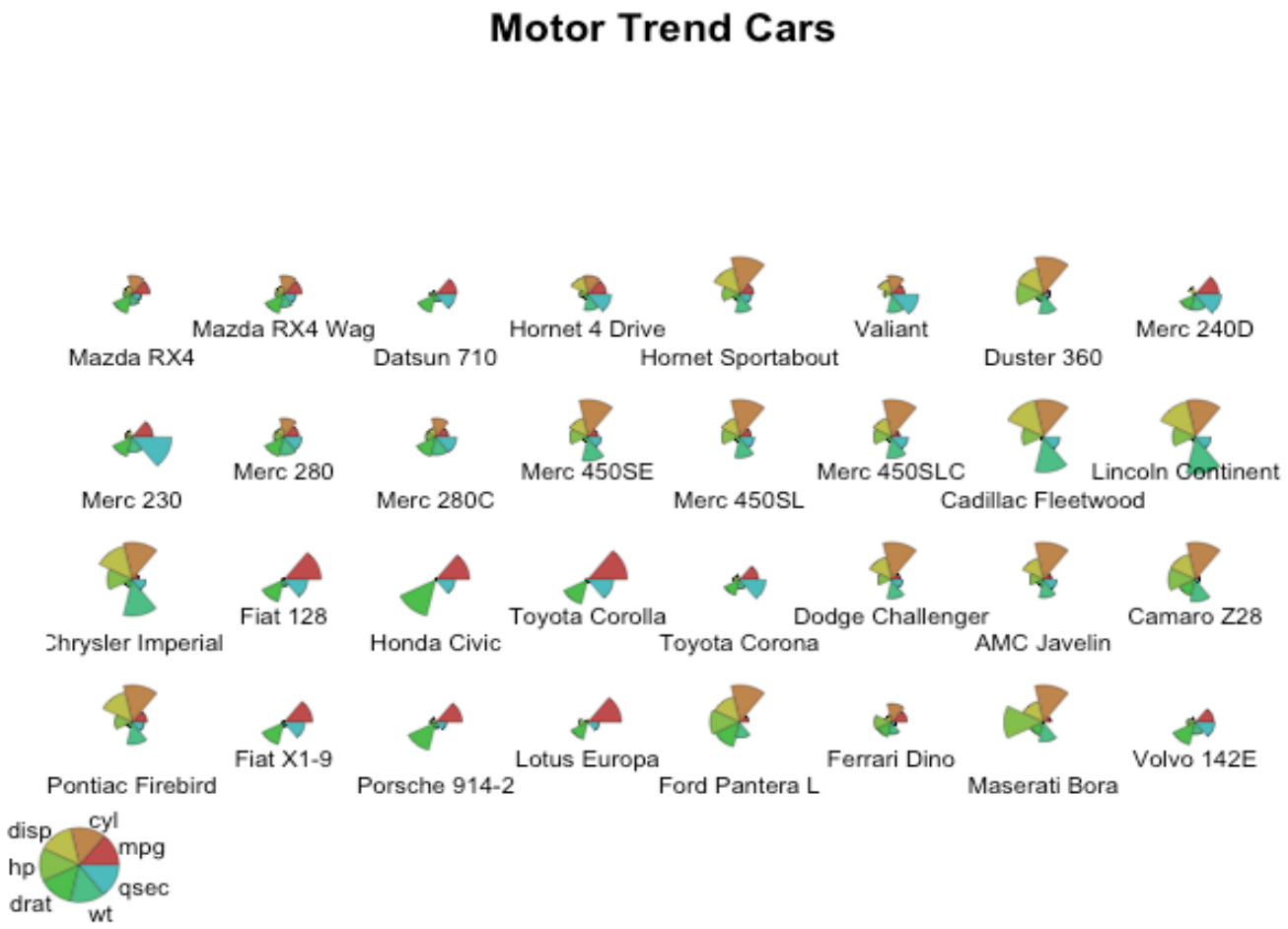
3D Scatterplot – Sepal length, Sepal width and Petal length



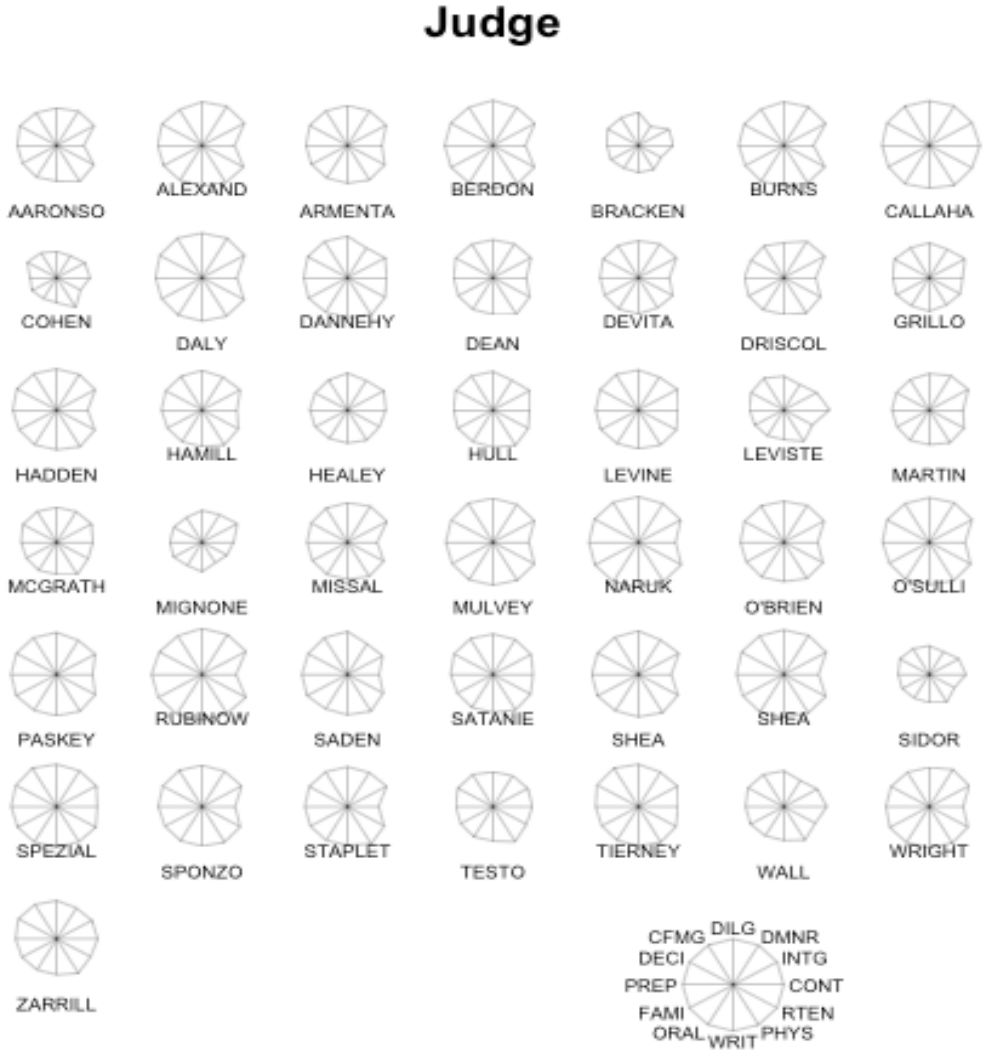
Growth Curves – Curves on Same Plot – Greenhouse Gas Emissions



Star Plot – MT Cars



Star Plot – Judges data



Chernoff Faces– The face of crime