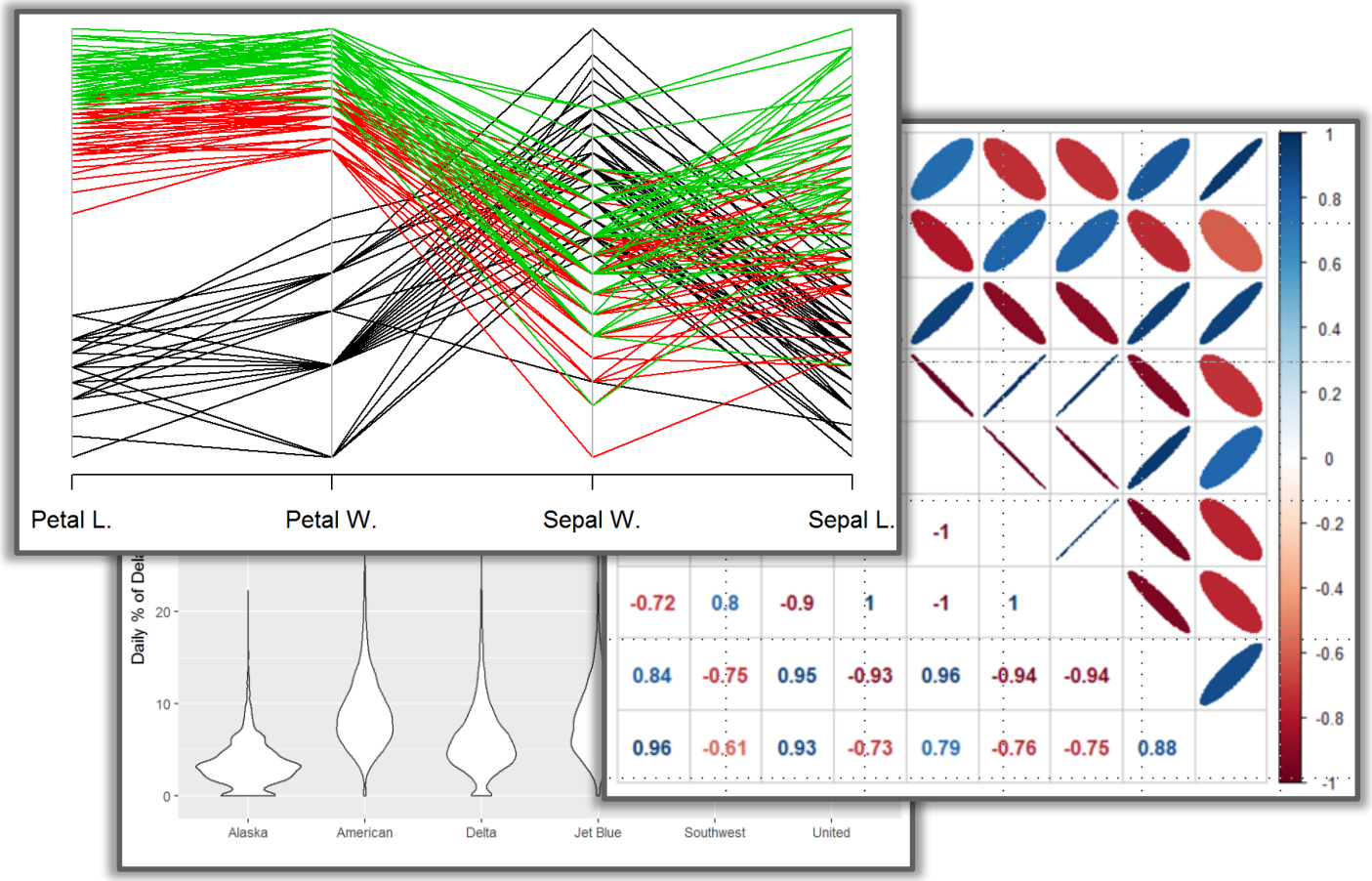
Exploratory Data Analysis: Overview



- Taxonomy of variable types
- Univariate analysis
- Bivariate analysis: relationships
- Multivariate analysis

Prof. Nagiza F. Samatova
samatova@csc.ncsu.edu
Department of Computer Science
North Carolina State University

NC STATE Executive Education

February 2019

Quantitative vs. Binary / Categorical Variables

- **Binary or Categorical: Can NOT be compared with <, >**
  - Binary: two values: sex:{F, M})
  - Polytomous: a finite set of values:
    - Nominal: cannot be compared: zip codes, country names

- **Quantitative Variable**
  - Continuous:
    - Have real numbers as values
    - Often represented as *floating point* variables
    - Examples: temperature, height, weight
  - Discrete:
    - Count: countably infinite set: number of traffic violations
  - Categorical Discrete
    - Polytomous Ordinal: a finite set of values that can be compared (<, >): (poor, good, excellent)

Quantitative vs. Binary / Categorical

Feature	Description	Example	Statistical Operation	Discrete vs. Continuous
Nominal	values are different names: provide enough info to distinguish one object from another (=, ≠)	zip codes, employee ID, eye colors, sex:{male, female}	**mode**, contingency, entropy, χ^2-test	discrete
Ordinal	values provide enough info to order objects (<, >)	grades (A, A-, B, B+) size (small, medium, large)	**median**, percentiles, rank correlation, run tests, sign tests	discrete
Interval	the differences between values are meaningful: allow ordering and subtraction but not other arithmetic operations	calendar dates, time, temperature in Celsius	**median, mean**, standard deviation, Pearson's correlation, t- and F-tests	both
Ratio	both differences and ratios are meaningful (*, /)	monetary quantities, counts, age, length, temperature	**mean, median, geometric mean, harmonic mean**, percent variation	continuous

EDA: Exploratory Data Analysis

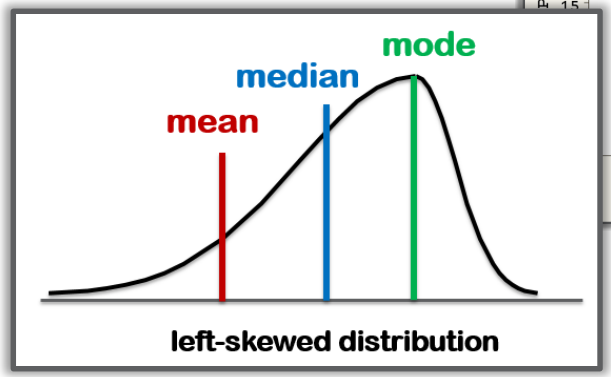
- **Univariate Analysis: one variable at a time**
  - Quantitative variable
  - Binary or categorical variable

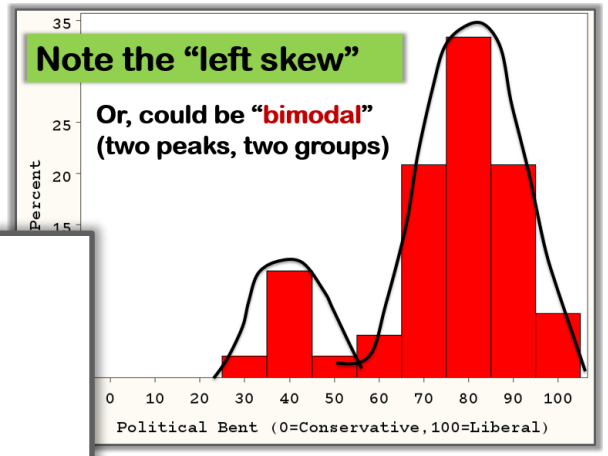
- **Bivariate Analysis: to explore relationships between two variables at a time**
  - Correlation between two quantitative variables
  - Association between two categorical or binary variables
  - By-group analysis: A quantitative variable within each group / category

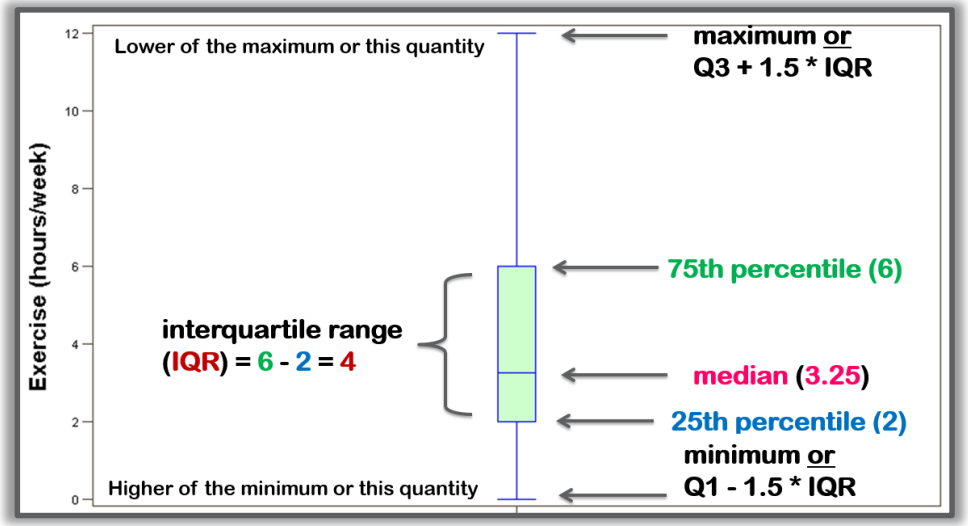
- **Multivariate Analysis: more than two variables at a time**
  - Visualization: Conditioning or facets

Univariate: Quantitative Var: Central Tendency & Variability

- Where is the center?
  - Central tendency
  - Arithmetic / geometric / harmonic mean, median, weighted mean/median, trimmed, mean, mode
- What is the variability or spread?
  - Variance (average distance from the mean)
  - Standard Deviation (sqrt (variance))
  - Range (max – min)
  - Percentile
  - Inter-Quartile Range (IQR)
- Data Distribution & Visualization
  - Shape, center, spread, left-skewed, right-skewed, outlier
  - Frequency table: bins with the missing values, outliers
  - Histogram
  - Density plots
  - Boxplots

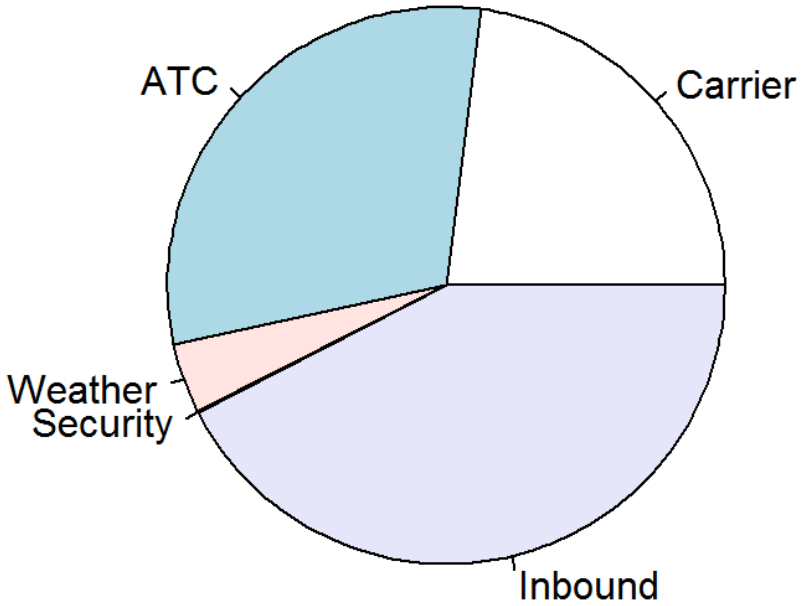


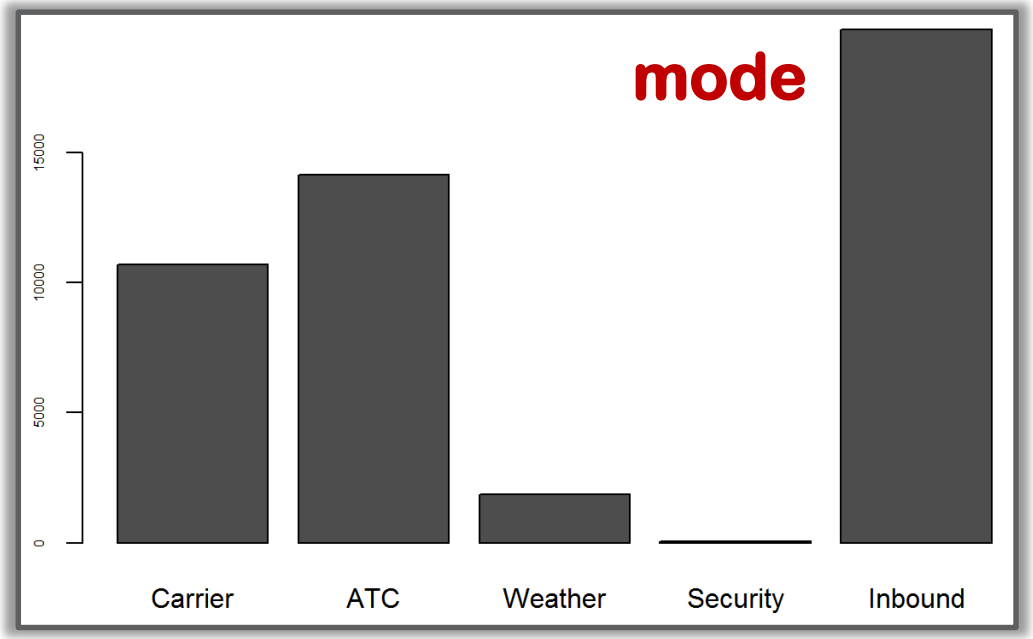




Univariate: Categorical Var: Mode & Expected Value

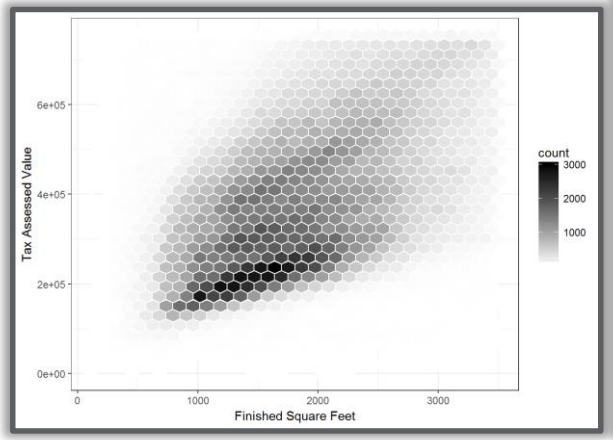
- **Mode**
  - the most commonly occurring category or value
- **Expected Value**
  - an average value of the numeric category based on category's probability of occurrence
- **Visualization**
  - Bar charts
  - Pie charts

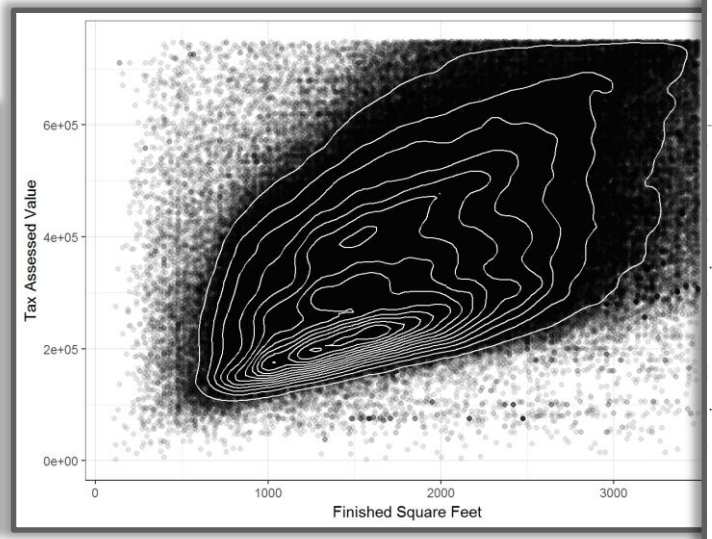


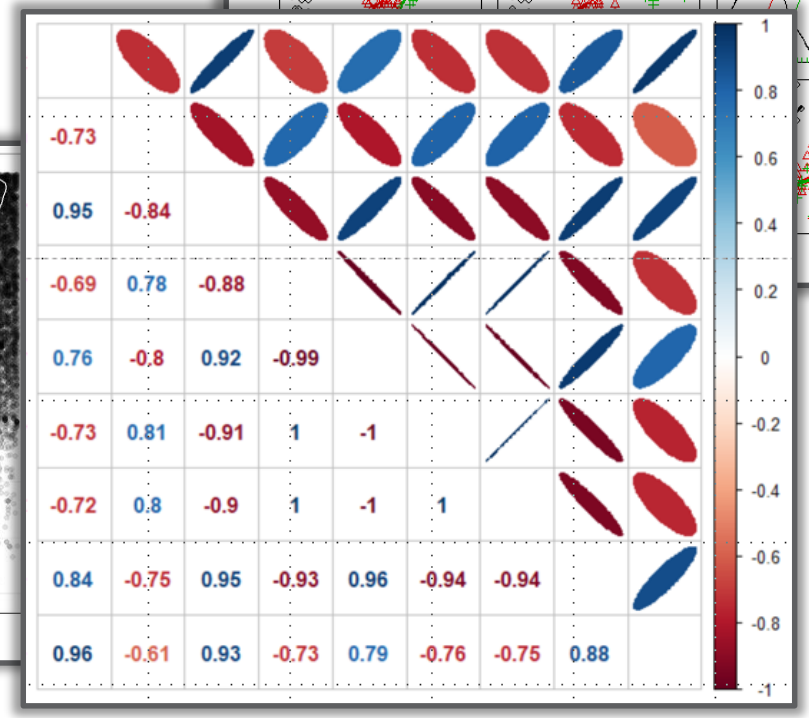


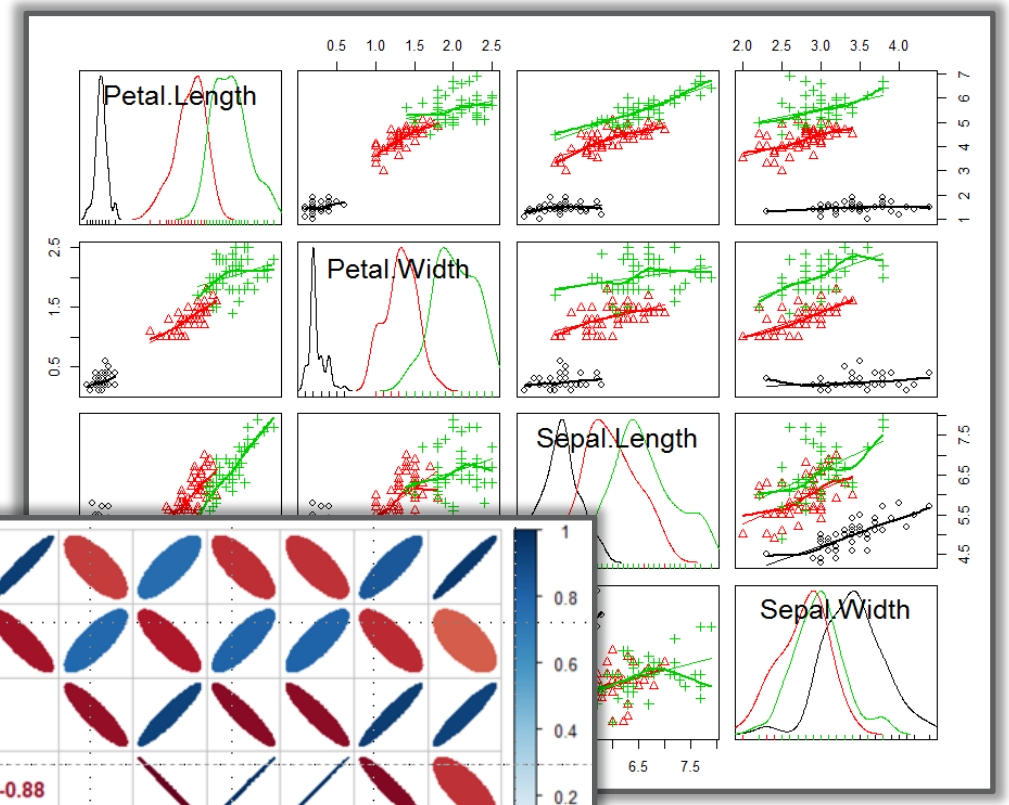
Bivariate: Quantitative Var: Correlation & Scatter

- **Correlation**
  - Among predictors
  - Between predictors and a target variable
- **Visualization**
  - Correlation matrix
  - Scatterplot matrix
  - Hexagonal binning
  - Contour plots





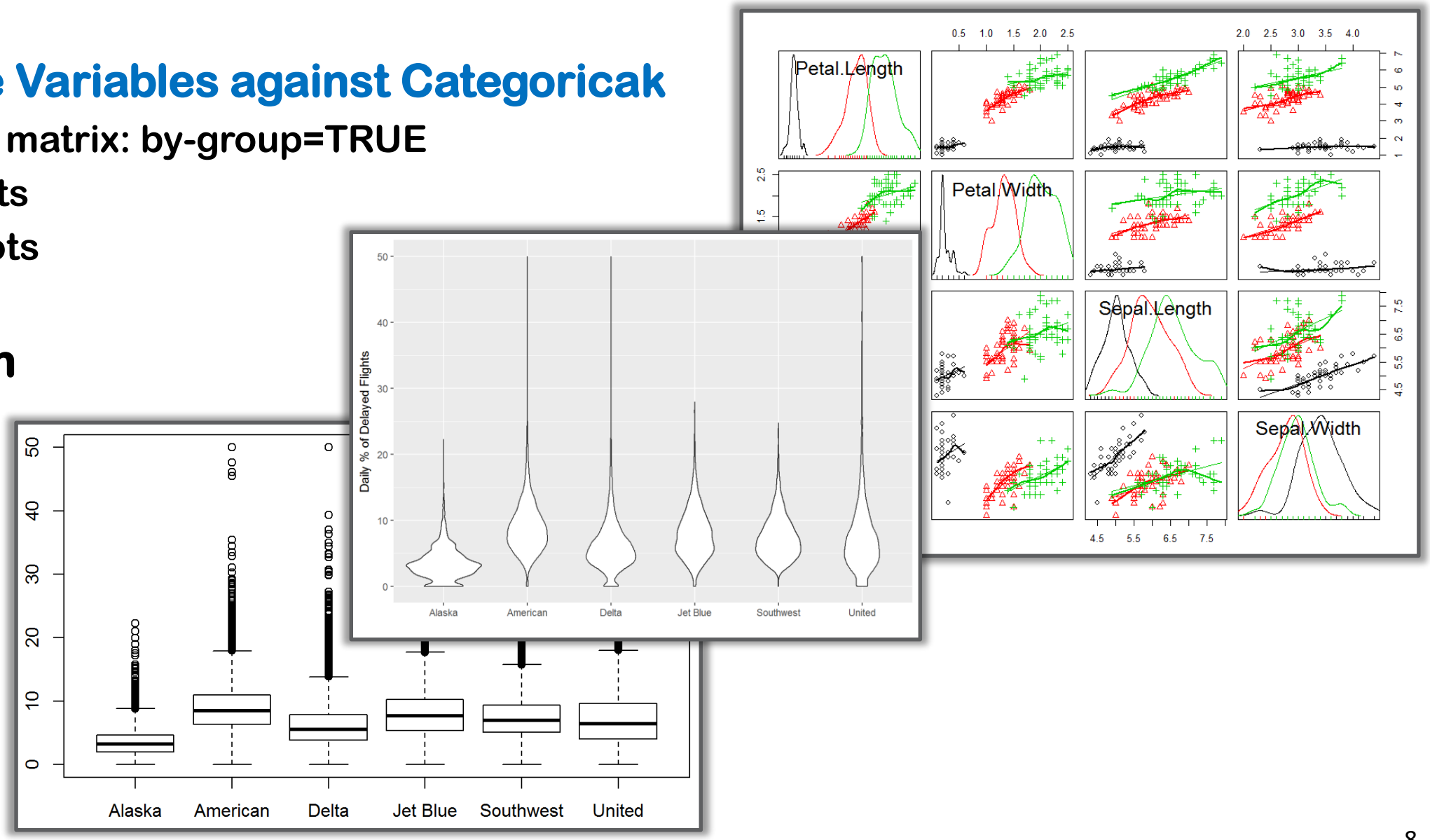




Bivariate: Quantitative against Categorical Var: Violin & Viz

- **Quantitative Variables against Categoricak**
  - Scatterplot matrix: by-group=TRUE
  - Density plots
  - Quartile plots

- **Visualization**
  - Box plots
  - Violin plots



Multivariate: Quantitative Var: Visualization

- Visualization
  - Conditioning or facets
  - Parallel coordinate plots
  - Cluster Dendrogram