CMPINF 2100

Introduction to Data Centric Computing

Week 10

Review: Key questions associated with choosing Plot Types

There are two main things you should consider when choosing a plot.

- First, are you studying a variable by itself (marginal) or a relationship between variables?

- Second, what are the data types of the variables you will include in the plot?

Studying a single variable – MARGINAL plots

- Marginal plots visualize the distribution of a single variable.
- Categorical or non-numeric variables are visualized with bar charts.
- Continuous or numeric variables are visualized with histograms, KDE plots, rug plots or composites of those together.
- The one caveat is that you should also consider the number of unique values because a numeric variable with just a few unique values can be explored as a categorical variable.

Studying the relationships between several variables

- Relationship plots are broken down by the types of relationships they are between.

- There are 3 types of relationships you need to consider:
  - Categorical-to-categorical relationships
  - Categorical-to-continuous relationships
  - Continuous-to-continuous relationships

Categorical-to-categorical relationships

- Categorical-to-categorical relationships show the counts of the combinations between variables.
- There are 3 main types you can use, but there are some considerations based on the number of variables you considering:
  - 2 variables: dodged bar charts, faceted bar charts, or heat maps
  - 3 variables: faceted dodged bar charts or faceted heat maps
  - 4 variables: column and row faceted dodged bar charts or column and row faceted heat maps
  - 5 variables or more: do not attempt

Categorical-to-continuous variable relationships

- Categorical-to-continuous relationships show how the distribution of the continuous variable changes across the categories (groups) of the categorical variable.

- The distribution is therefore conditioned on a given category.

- The conditional distribution can be represented and summarized in different ways.

Categorical-to-continuous variable plots: 1 categorical and 1 continuous variable

- **CONDITIONAL KDE** - color the KDE of the continuous variable by the categories of the categorical variable - must remove the sample size effect
  - relationship is represented by the separate colored distributions
  - shows if the distributional shape varies across the categories
  - shows if the most frequent value (the mode, value with the highest density) changes across categories
  - does not show important summary statistics
  - best suited for a small number of categories
- **VIOLIN PLOT** - conditional distribution represented by separate filled KDEs oriented vertically like violins
  - relationship represented typically with the continuous variable on the y-axis and the categorical variable on the x-axis
  - shows the conditional distributional shape
  - can include summary statistics
  - can be difficult to look at if there are many categories

Categorical-to-continuous variable plots: 1 categorical and 1 continuous variable

- **BOXPLOT** - summary statistics of the continuous variable given the categories of the categorical variable
    - relationship represented typically with the continuous variable on the y-axis and the categorical variable on the x-axis
    - shows if the median changes across the categories
    - shows if the variation of the continuous variable changes across the categories due to the height of the box and length of the whiskers
    - separation of the boxes is a strong indicator that the distribution of the continuous variable depends on the categorical variable
    - cannot show the distributional shape
    - scales well to many categories
- **POINT PLOT** - focus on the mean of the continuous variable given the categories of the categorical variable
    - relationship represented typically with the continuous variable on the y-axis and the categorical variable on the x-axis
    - only shows the conditional averages but includes the confidence interval on the average
    - separation of the confidence interval represents the average does vary across groups
    - does not answer any other question beside changes in the mean

Categorical-to-continuous variable plots: 2 categoricals and 1 continuous variable

- **Faceted conditional KDE** - color the KDE of the continuous variable by a categorical variable and facet by a second categorical variable
- **Boxplot** - 2 options
  - facet the boxplot by a second categorical variable
  - color (hue in Seaborn) by a second categorical variable
- **Violin plot** - 2 options
  - facet the boxplot by a second categorical variable
  - color (hue in Seaborn) by a second categorical variable
- **Point plot** - 2 options
  - facet the boxplot by a second categorical variable
  - color (hue in Seaborn) by a second categorical variable

Continuous-to-continuous variable relationships

- Continuous-to-continuous relationships show how one continuous variable relates to another continuous variable.
- The plot types are based on whether the "raw" data are shown or if summary statistics are used to represent the relationship.
- Most of these plots can be colored by a third variable and faceted by categorical variables.

Continuous-to-continuous plot types

- **SCATTER PLOT**- shows the raw observations between the two variables
    - can be colored by a categorical variable to show if the relationships depend on categories
    - can be colored by a continuous variable to show if a third continuous variable impacts the relationship between the two
    - can be faceted by categorical variables to show if the relationships depend on categories
    - can combine color and facets to include the impact of 3 categorical variables, but this approach is best suited when the categorical variables have less than 5 categories
- **TREND PLOT** - includes a "best fit line" to show the general linear relationship between the two variables
    - Raw data shown as markers and the trend line is shown on top of the markers
    - can be colored by a categorical variable to show if the relationships depend on categories
    - can be colored by a continuous variable to show if a third continuous variable impacts the relationship between the two
    - can be faceted by categorical variables to show if the relationships depend on categories
    - can combine color and facets to include the impact of 3 categorical variables, but this approach is best suited when the categorical variables have less than 5 categories

Continuous-to-continuous plot types

- **<u>CORRELATION PLOT</u>** - does not show the raw data instead shows the correlation coefficient to summarize the linear relationship as a heatmap
    - A cell in the heatmap corresponds to the correlation coefficient between a pair of variables. Different cells show the correlation coefficient between different pairs.
    - Is symmetric around the main diagonal. The upper triangle “above” the main diagonal shows the same values as the lower triangle “below” the main diagonal.
    - The main diagonal simply shows each variable is perfectly correlated with itself and thus the main diagonal ALWAYS shows correlation coefficient values of 1.
    - Must use a diverging color palette with lower bound of -1, upper bound of 1, and a midpoint (center) of 0.
    - Scales to many pairs of variables and thus can show the linear relationship between many continuous variables.
    - Can be faceted by a categorical variable to show if the linear relationships change across categories.

Continuous-to-continuous specialized plot type

- **PAIR PLOT** – combines marginal and relationship plot types
  - Main diagonal shows the marginal distribution of each continuous variable
  - Off-diagonal shows the relationship between each pair of variables.
    - The relationship is typically represented as a scatter plot
  - Provides more detail than a CORRELATION PLOT because the raw data are shown rather than a summary statistic.
  - Does NOT scale well to many variables because there are simply too many subplots generated.
    - Best suited for less than 10 to 12 numeric variables.
  - Can color by a categorical variable but the sample size effect MUST be removed for the main diagonal conditional KDE plots.