Methods in Open Science

Exploring Data

Image from NASA: https://commons.wikimedia.org/wiki/File:Earth_Western_Hemisphere_transparent_background.png#filelinks

WV VIEW

AmericaView
Empowering Earth Observation Education
AmericaView.org - Est. 2003

In this module, we return to working with data. Specifically, we will discuss how data can be explored, visualized, summarize, cleaned, and prepared.

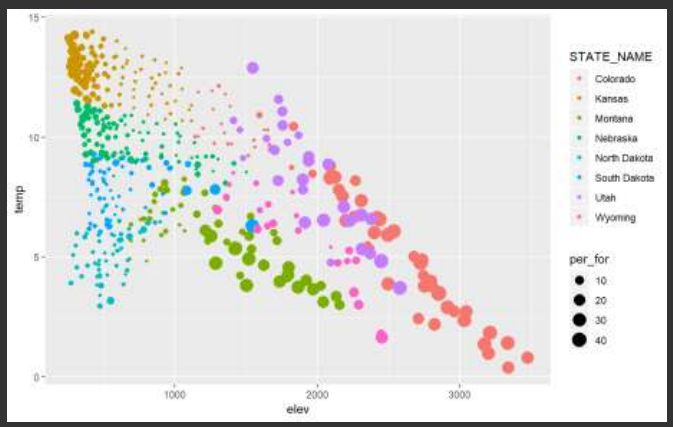
Visualizing Data

2

We will start off with visualizing data using graphs.

Graphs

- Summarize variables
- Explore relationships between variables
- Map variables to graphical parameters



3

In order to visualize data graphically, we can map specific variables to graphical parameters such as position, size, color, shape, etc. A major component of generating graphs is determining how best to represent our data visually. This is the primary focus of this section. We will then turn our attention to specific types of graphs.

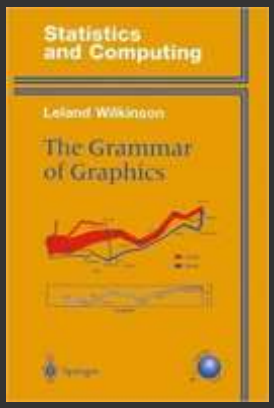
Note that we will not talk about specific graphing and data visualization tools here. Those topics are reserved for the Python and R material. Within Python, I demonstrate the matplotlib library and other associated libraries (e.g., seaborn and pandas). In the R material, I focus on ggplot2.

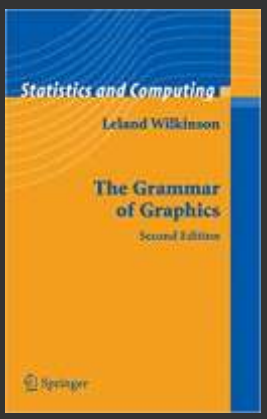
Book Recommendation

- Leland Wilkinson
- 1st Edition: 1999
- 2nd Edition: 2005
- Grammar/philosophy for visualizing data
- Map data to graphical parameters (aesthetic mappings)

https://www.springer.com/gp/book/9780387245447

https://towardsdatascience.com/a-comprehensive-guide-to-the-grammar-of-graphics-for-effective-visualization-of-multi-dimensional-1f92b4ed4149b





4

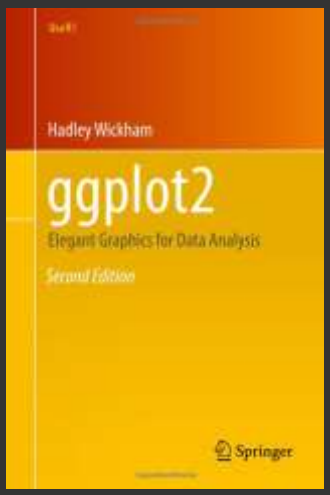
The *Grammar of Graphics* was proposed by Leland Wilkinson and offers a framework, philosophy, and grammar for visualizing data. A central component of this philosophy is the assignment of data to graphical parameters or defining aesthetic mappings. The *Grammar of Graphics* has greatly informed effective data visualization and graph design and is also useful for making maps.

I use this framework in this section to guide the discussion. In the R material, you will see that the ggplot2 package created by Hadley Wickham implements this philosophy.

Book Recommendation

Wickham, H., 2016. *ggplot2: elegant graphics for data analysis*. springer.

https://hadley.nz/
https://ggplot2-book.org/

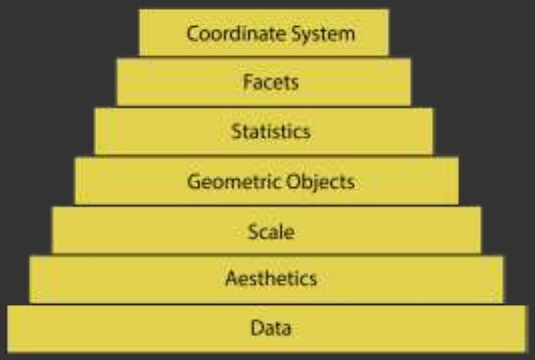


5

This is a great text that explains the graphing functionality of ggplot2, which is based on the *Grammar of Graphics*. Again, this R package is explored in the R material associated with this course.

Grammar of Graphics

- Data = Always start with the data, identify the dimensions you want to visualize.
- Aesthetics = Confirm the axes based on the data dimensions, positions of various data points in the plot. Also check if any form of encoding is needed including size, shape, color, and so on which are useful for plotting multiple data dimensions
- Scale = Do we need to scale the potential values and/or use a specific scale to represent multiple values or a range?
- Geometric objects = This would cover the way we would depict the data points on the visualization. Should they be points, bars, lines, and so on?
- Statistics = Do we need to show some statistical measures in the visualization like measures of central tendency, spread, and/or confidence intervals?
- Facets = Do we need to create subplots based on specific data dimensions?
- Coordinate System = What kind of a coordinate system should the visualization be based on?



https://www.springer.com/gp/book/9780387245447

https://towardsdatascience.com/a-comprehensive-guide-to-the-grammar-of-graphics-for-effective-visualization-of-multi-dimensional-1f92b4ed4149b

6

Based on the *Grammar of Graphics*, all graphs begin with data. The data scientist must decide what dimensions of the dataset should be visualized in the graph space.

Aesthetic mappings are then used to visualize these data based on their location in the graph space or their assignment or visualization with a graphical property, such as color, shape, size, etc. The appropriateness of different graphical parameters for visualizing data depends on the data in questions (i.e., nominal, ordinal, or numeric).

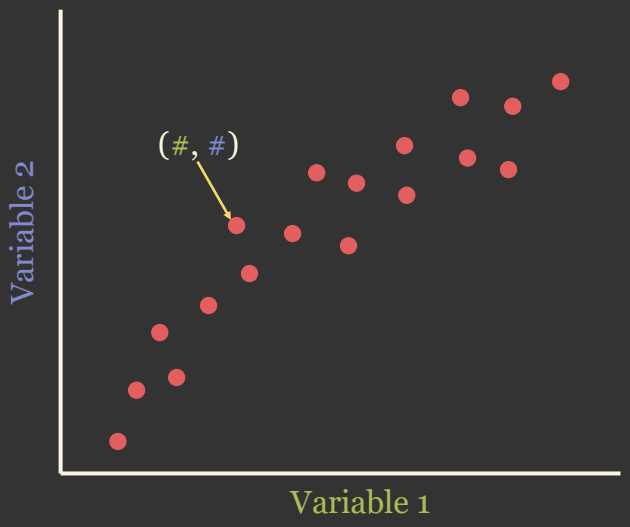
Next, appropriate scales must be defined. This could include ranges in the x- and y-directions, appropriate color ramps and binning methods, size ranges, sets of shapes, etc. You must also decide what geometric objects should be used to represent data. Options include points, bars, such as bar graphs, or lines, such as time series graphs.

We can also include data summarizations and statistics, such as measures of central tendency and/or variability, and confidence intervals. Facets can be used to break the data into multiple plots or maps based on a categorical or grouping variable. Lastly, coordinate systems can be altered. For example, graphs can make use of Cartesian coordinate space or polar coordinates.

Position

- For all levels of measurement
- Appropriate graph types will depend on level of measurement of variables and relationships being explored

What if you want to show more than two variables?



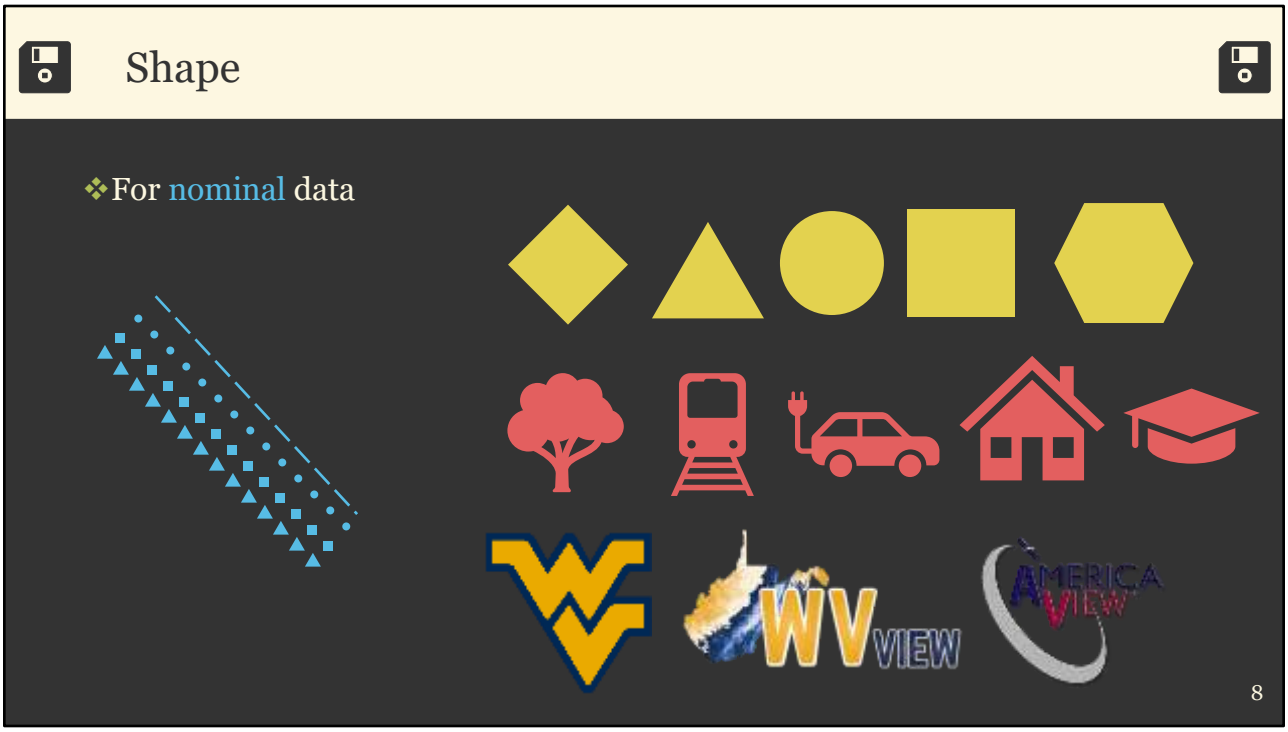
7

We will now discuss how different graphical parameters can be used to effectively display data. We will begin with position.

In a graph space, variables can be used to define the position of symbols relative to axes. Most graphs use a two-dimensional Cartesian space with an x- and a y-axis. When designing maps, positions can be defined using latitude and longitude or map projection coordinates.

As we will discuss in more detail later in this section, mapping a variable to the x-axis and another to the y-axis allows us to visualize the relationship between the two variables.

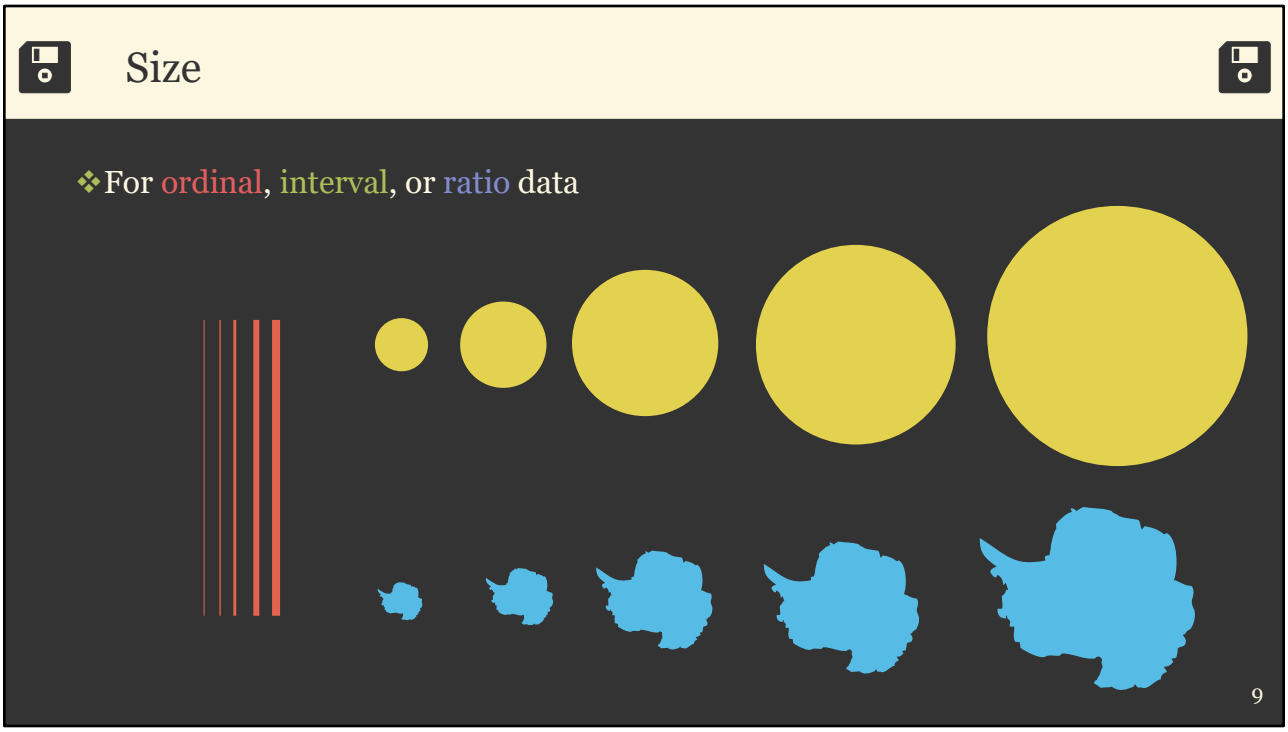
Note that some graphs make use of other spaces as opposed to two-dimensional Cartesian space. For example, it is possible to include a third, or z, dimension to map points in 3D space. This can be used to compare the relationship between three variables by plotting points within the volume of a cube. Some graphs make use of polar coordinates, such a pie charts. For maps, a variety of coordinate systems are defined as different map projections.



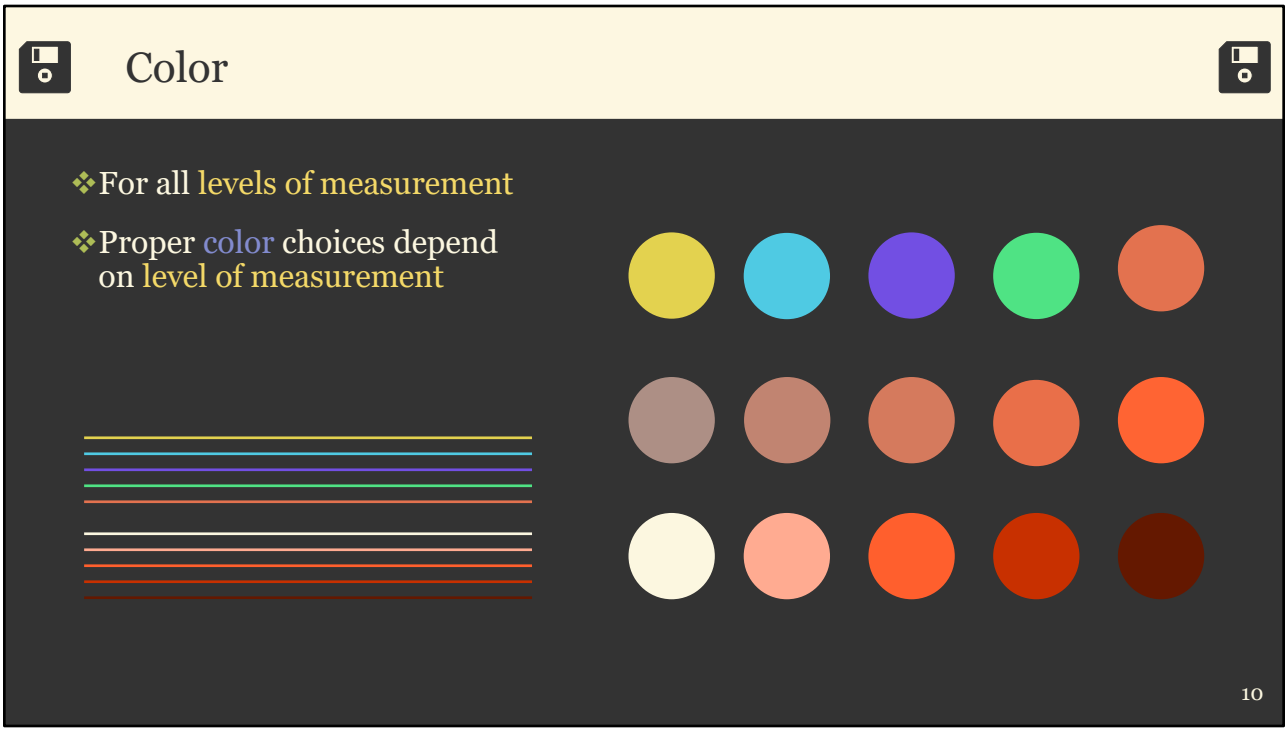
Shape is only appropriate for visualizing nominal variables since there is no implied increasing or decreasing order. For example, how could you represent crime rates for cities using shape? This really wouldn't make sense.

Different types of shape-based symbols can be used. Geometric relies on common, generally simple geometric shapes, such as circles, squares, rectangles, ovals, diamonds, stars, and hexagons. Mimetic symbols mimic the feature they are representing, such as the examples provide on the slide. Lastly, pictorial symbols use pictures, such as company logos.

It is also possible to categorize line features with shapes using repeating geometric shapes along the line segments.



Size should only be used to represent ordinal or numeric data since there is an implied order or sequence from small to large. Larger sizes will indicate larger quantities while smaller sizes will indicate smaller quantities. For lines, the line width or thickness equates to size and again is only appropriate for numeric data. Examples of maps that use size include proportional symbol, graduated symbol, and cartograms.



Different aspects of color can be used to represent different types of data.

For example, unordered colors that change by hue (or change by base color) can be used to represent nominal data, as long as no order is implied by the color choices. The top set of circles and top set of lines are examples.

To use color to represent ordinal or numeric data, an order should be implied. This generally involves increasing the saturation and/or lightness of a base hue, as demonstrated by the middle-set of circles (where saturation is varied) and the bottom set of circles and lines (where lightness is varied).

It is generally argued the color is overused. Part of the issue is that humans are generally not great at relating changes in saturation or lightness to a change in the magnitude of a quantity. However, color can be effective if used correctly and thoughtfully.

Spacing

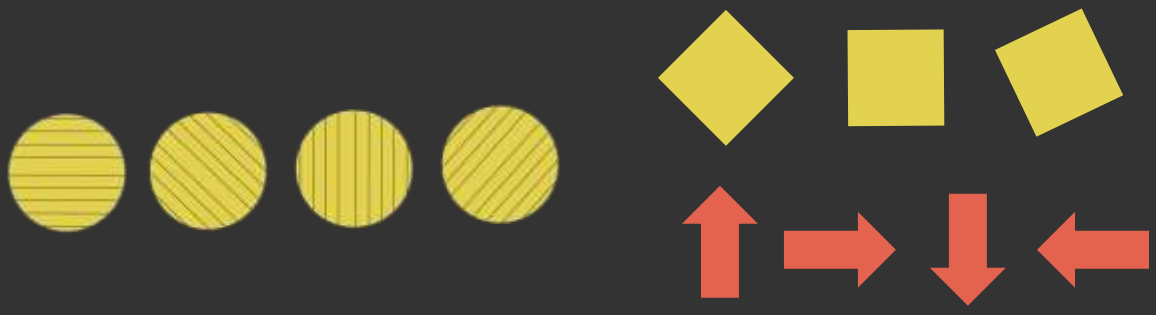
- For ordinal, interval, or ratio data

11

Spacing can be used to represent ordinal or numeric variables but is not appropriate for nominal data since there is an implied ordering. Generally, denser spacing indicates a larger quantity whereas less dense spacing represents a smaller quantity. For lines, the spacing is represented by the length and spacing between dashes. Areal features can be filled with a pattern that varies only by density.

Orientation

- For nominal data

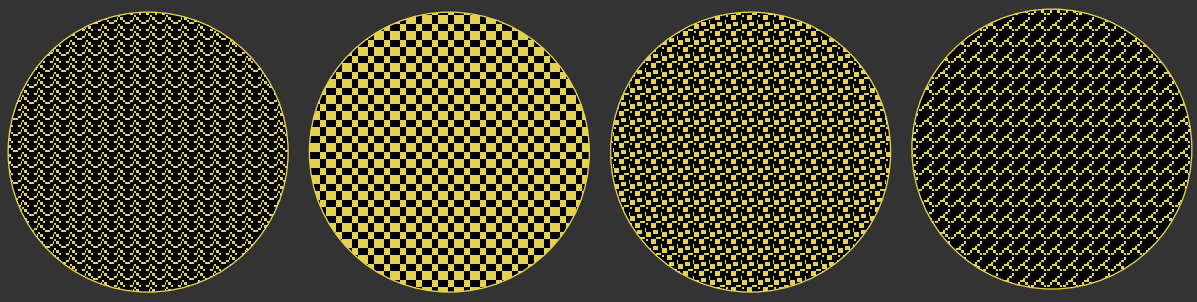


12

Orientation can be used to represent nominal data but is inappropriate for numeric data, since there is no implied order. To use orientation the same symbol is used to represent a point or fill an areal feature, with only the orientation of the pattern changed.

Arrangement

- For qualitative data



13

Another option is to use arrangement to represent nominal data. In this case, different base patterns with no implied ordering are used.

Comparisons

Graphical Parameter	Nominal	Ordinal	Numeric (Interval and Ratio)
Shape	Good	Poor	Poor
Size	Poor	Marginal	Marginal
Color (Hue)	Good	Good	Marginal
Color (Saturation)	Poor	Marginal	Marginal
Color (Lightness)	Poor	Good	Marginal
Spacing	Poor	Marginal	Marginal
Orientation	Good	Poor	Poor
Arrangement	Good	Poor	Poor

Slocum, T.A., McMaster, R.B., Kessler, F.C. and Howard, H.H., 2014. *Thematic cartography and geovisualization*. CRC Press.

14

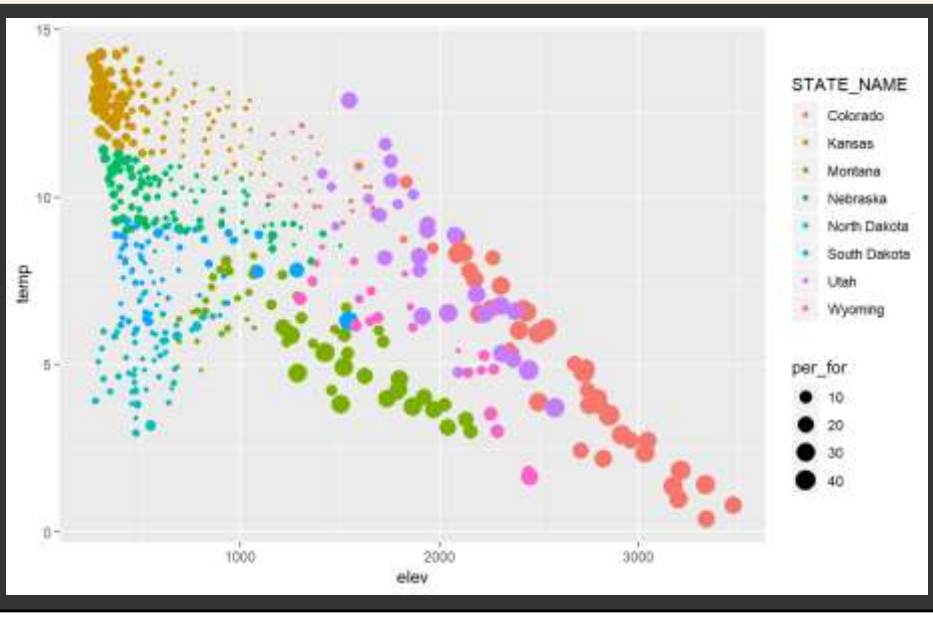
This table was modified from one presented in a textbook focused on map design and geovisualization. It summarizes what graphical parameters are appropriate or best to use for different levels of measurement. Note that numeric includes both interval and ratio data.

Shape is only appropriate for nominal data since there is no implied ordering. Size, since there is implied ordering, is not appropriate for nominal data but can be used for ordinal or numeric data, since an implied ordering is desired.

Hue works well for nominal data and can be used for ordinal data if hues are selected to represent some implied ordering. Generally, altering only the hue is avoided for representing numeric data. The color saturation and/or lightness can be used for ordinal and numeric data but are inappropriate for nominal data because an ordering is implied.

Spacing can be used for ordinal and numeric data while orientation and arrangement are only applicable to nominal data.

Graphs and Aesthetic Mappings



15

Let's now explore the aesthetic mappings used in this graph. These data represent county-level means for all counties occurring in the High Plains regions of the United States.

Elevation, a numeric variable, is mapped to the x-axis position while temperature, also a numeric variable, is mapped to the y-axis position. This allows for the relationship between the two variable to be visualized. Generally, it appears that counties that occur at a higher mean elevation tend to have a lower average mean annual temperature.

The size of the point symbol is mapped to another continuous variable: percent of the county land area that is forest. Generally, it appears that there might be some correlation here. Counties at higher elevations appear to generally have a larger percentage of forest cover.

Lastly, an unordered set of hues is used to differentiate the points, which represent counties, by the state in which they occur.

Note that I could have chosen to use different mappings. For example, I could have used changing saturation or lightness to represent percent forest cover and different shapes or point symbols to represent the state in which the county occurs. It is generally a good idea to experiment with different aesthetic

mappings to assess the effectiveness for specific data and use cases.

Summary of Key Points

- The *Grammar of Graphics* offers a guide to effectively and appropriately visualizing data using graphical parameters or aesthetic mappings.
- What graphical parameter (e.g., shape, size, color, spacing, orientation, and arrangement) is best to represent a specific variable?
  - Often depends on the variables level of measurement
- Properly designed graphs take the *Grammar of Graphics* into account to create effective and informative data visualizations.

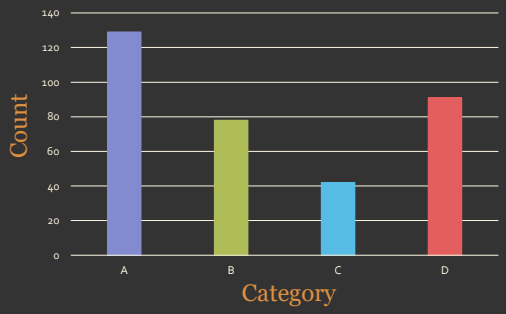
Describing and Summarizing Data

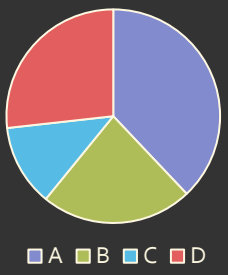
17

We will now discuss means to describe and summarize data using methods appropriate or meaningful for different data types.

Summarizing Categorical Data

- How many categories?
- Are categories nested?
- Number of occurrences by category
- Percentage of total observations within each category





Category	Count	Percentage
A	129	37.9%
B	78	22.9%
C	42	12.4%
D	91	26.8%

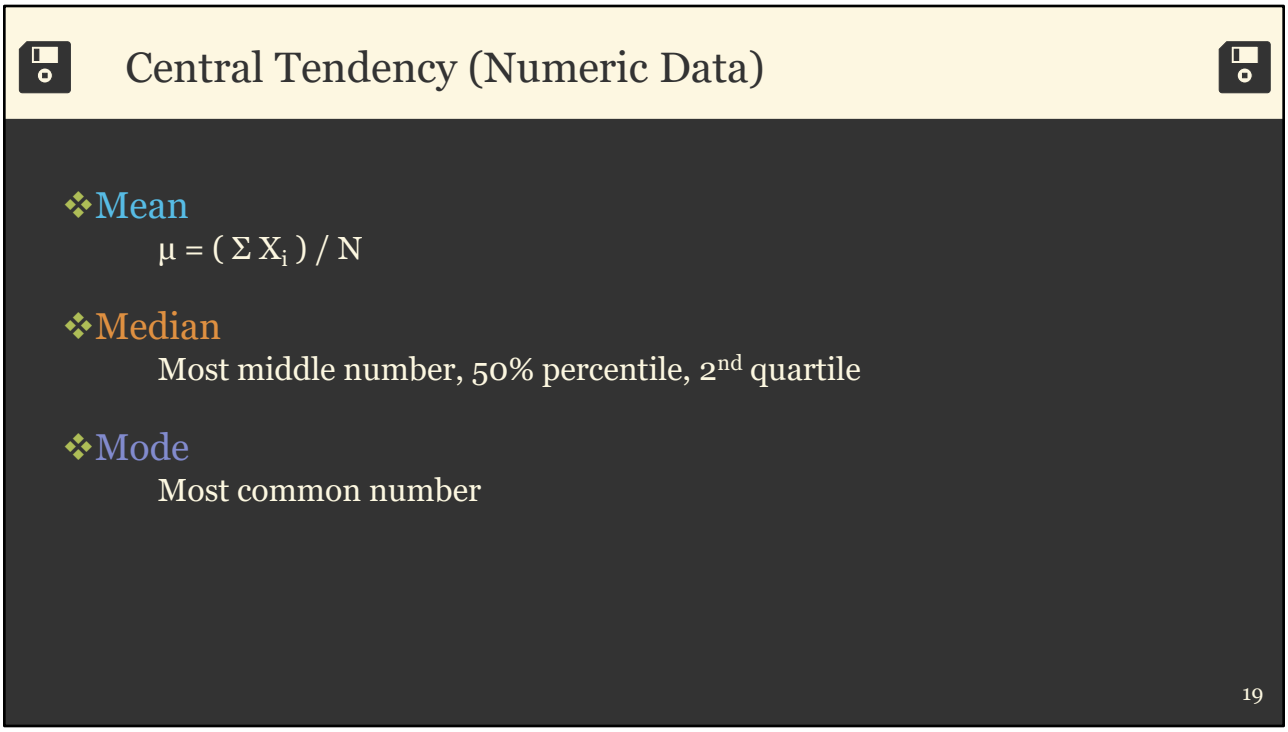
18

For nominal and ordinal data, we may be interested in the number of features that occur in each category or the proportion or percentages of features by category. For example, using the charts in this slide, we may want to count the number of counties by state or the percentage of the counties occurring in each state.

Note that sometimes categories are nested. For example, we could describe the surface cover within an extent as either water, forest, grass, barren, or developed. Forest could be further subdivided into forest types, such as deciduous, evergreen, and mixed. Developed could be differentiated into residential, commercial, and industrial types.

Bar charts are generally preferred for displaying counts by categories as opposed to pie charts since humans are better at judging relative lengths as opposed to relative sizes. Generally, you should avoid using pie charts.

Sometimes, categorical data and associated counts or percentages can better be summarized in a table as opposed to with a graph.



For numeric data, we are generally interested in understanding the central tendency and the variability in the data. Central tendency relates to the most common, most middle, or most frequent values whereas variability relates to how distributed or dispersed values are within a set of numbers.

Measures of central tendency include the mean (sum up all values and divide by the number of records), the median (the most middle value; 50% of the values are below this value and 50% are above), and the mode (the most commonly occurring value).

For example, you could calculate the average or mean of exam scores by adding up all exam scores and dividing by the number of exams.

One key issue is that central tendency only describes one aspect of the data. For example, two sets of exam scores could have the same mean or median; however, the range of test scores could be very different. For example, a B average could be obtained by earning Bs on all exams or by earning a more variable mix of scores, such as a mix of Cs and As. As a result, it is important to consider both the central tendency and variability.

When is it more appropriate to use the median as opposed to the mean? This can be a tricky issue, and many practitioners would disagree on this point. One

issue with the mean is that it tends to be more heavily impacted by extremely large or small values in comparison to the median. So, if you don't want large or small values to have a large effect on your measure of central tendency, the median may be more appropriate than the mean. We will discuss this in more detail in the section on statistical inference.

Variability (Numeric Data)

- **Range**
  Maximum – Minimum
- **Population Variance**
  $\sigma^2 = \Sigma (X_i - \mu)^2 / N$
- **Population Standard Deviation**
  $\sigma = \text{sqrt}[\Sigma (X_i - \mu)^2 / N]$

20

This slide describes measures of variability for numeric data. The range is simply the maximum reported value minus the minimum. This measure tends to be heavily impacted by very large of very small values since it only considers the largest and smallest values. For example, If 20 students in a course took an exam and all earned grades above 70% except for one student, who earned a grade of 15%, this low grade would have a large impact on the resulting range that may make the measure less informative.

Variance is calculated by subtracting the mean value from each value, squaring the differences, summing the squared difference, then diving by the number of samples. Since the difference is squared, the units of variance will be in the square of the units of the variable of interest. For example, if you are summarizing length measurements in meters, the variance would be in units of square meters. The standard deviation is simply the square root of the variance. Taking the square root will return the original unit of measurement.

Example Data: Seeds

area	perm	compactness	lkernel	wkernel	asymn	lgroove	type
Min. :10.59	Min. :12.41	Min. :0.8081	Min. :4.899	Min. :2.630	Min. :0.7651	Min. :4.519	1:70
1st Qu.:12.27	1st Qu.:13.45	1st Qu.:0.8569	1st Qu.:5.262	1st Qu.:2.944	1st Qu.:2.5615	1st Qu.:5.045	2:70
Median :14.36	Median :14.32	Median :0.8734	Median :5.524	Median :3.237	Median :3.5990	Median :5.223	3:70
Mean :14.85	Mean :14.56	Mean :0.8710	Mean :5.629	Mean :3.259	Mean :3.7002	Mean :5.408	
3rd Qu.:17.30	3rd Qu.:15.71	3rd Qu.:0.8878	3rd Qu.:5.980	3rd Qu.:3.562	3rd Qu.:4.7687	3rd Qu.:5.877	
Max. :21.18	Max. :17.25	Max. :0.9183	Max. :6.675	Max. :4.033	Max. :8.4560	Max. :6.550	

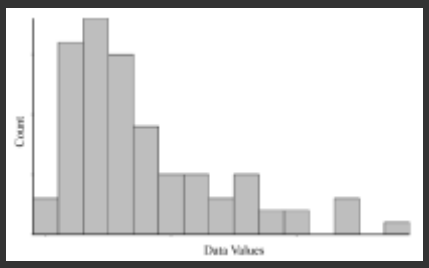
21

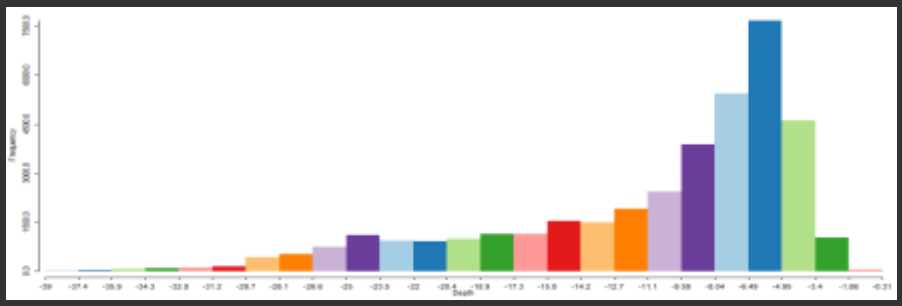
This table was generated using the R software and represents a summarization of the variables associated with the Seeds dataset. Note that all numeric data are summarized using the following measures: minimum, 1st quartile, median, mean, 3rd quartile, and maximum. We will discuss quartiles in a later slide.

For the nominal "type" field, the count of each unique type is provided. Here the codes 1, 2, and 3, are stand-ins for the three different varieties of wheat kernels: Kama, Rosa, and Canadian, respectively. So, this is an example of a number that should be treated as nominal data.

Histograms

- Univariate graph (shows one variable)
- x-axis: attribute or quantity of interest
- y-axis: frequency or count





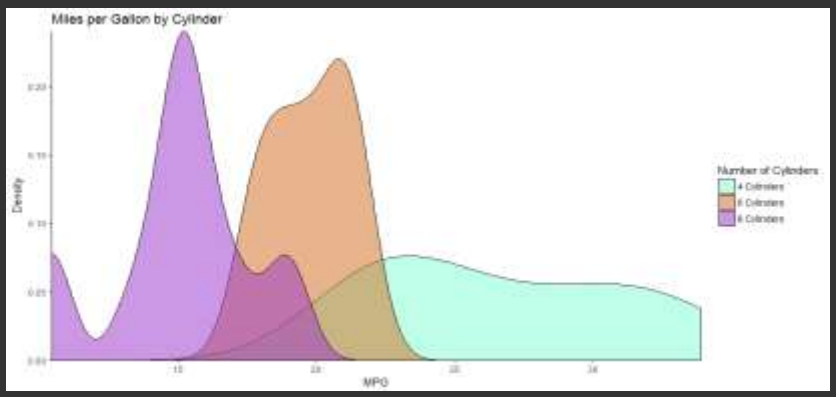
22

Histograms are one means to explore the distribution of a single variable and are an example of a univariate graph (i.e., a graph that only shows one variable). The x-axis represents the attribute or quantity of interest, and the y-axis will be the frequency or count. Data ranges are binned and the number of samples within each data range are counted. This allows us to visualize the distribution of the data to assess central tendency, dispersion, skewness, or the occurrence of outliers.

How the data are binned can have a large impact on how patterns are represented. Smaller bins tend to highlight more local patterns whereas larger bins will represent a more generalized pattern. There is not necessarily a correct bin width. This depends on the patterns of interest and the data being graphed.

Kernel Density Plots

- Univariate graph (shows one variable)
- Peaks = more common or frequent values
- Troughs = less common or frequent values



23

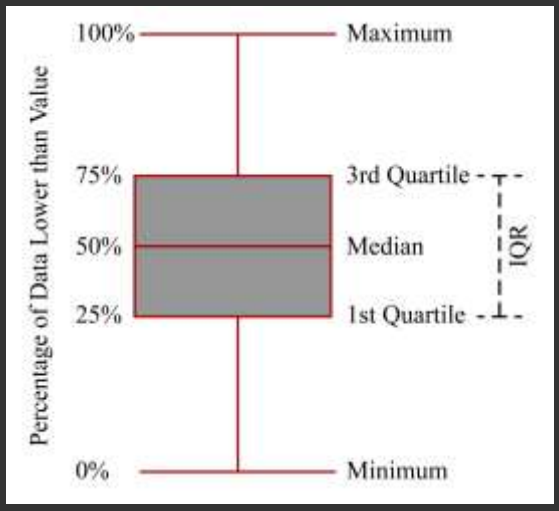
Similar to histograms, kernel density plots are used to show the distribution of a single variable (i.e., are univariate graphs). The x-axis will be the variable of interest and the y-axis will be density, count, or frequency. The curves are created using a kernel density function. Peaks indicate common values and troughs indicate less frequent values.

You can also use these graphs to compare data distributions for different categories, as shown in the example graph. We will discuss these types of comparisons later. Since histograms and kernel density plots are showing the same patterns and can have the same axes, they can be plotted in the same graph space.

The kernel density function can be adjusted to show more local or generalized patterns, similar to changing the bin width for histograms.

Box Plots

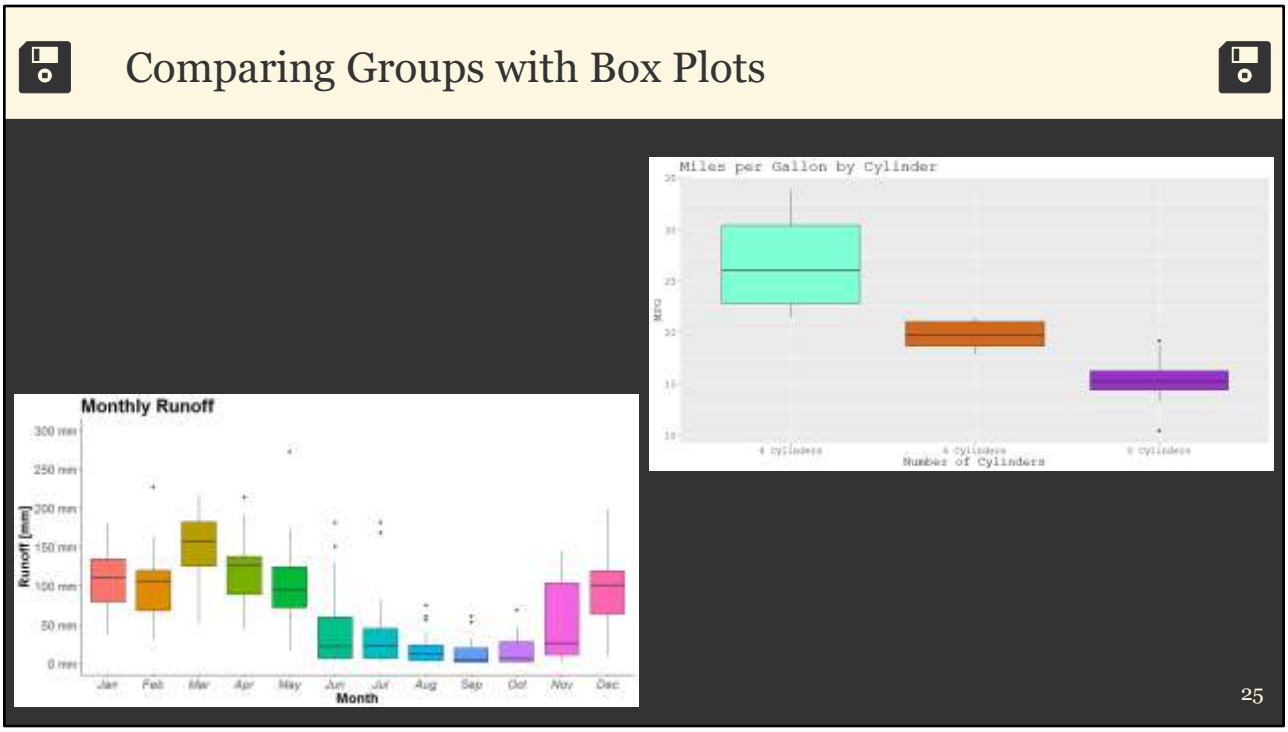
- Minimum = lowest value
- Maximum = highest value
- 1st Quartile = 25% of data below, 75% above
- 3rd Quartile = 75% of data below, 25% above
- Interquartile Range (IQR) = range of values between 1st and 3rd quartile



24

A boxplot is used to summarize the distribution of a numeric variable. It shows the lowest value (minimum), highest value (maximum), and median. Note that the mean is not commonly plotted on box plots.

The 1st quartile is the value that has 25% of the values below it and 75% above. In contrast, the 3rd quartile has 75% of the values below and 25% above. The range of values between the 1st and 3rd quartile is called the interquartile range, or IQR. Note that the median is the 2nd quartile (50% of the values are below it and 50% are above it). The maximum would be the 4th quartile (100% of the values are below it).



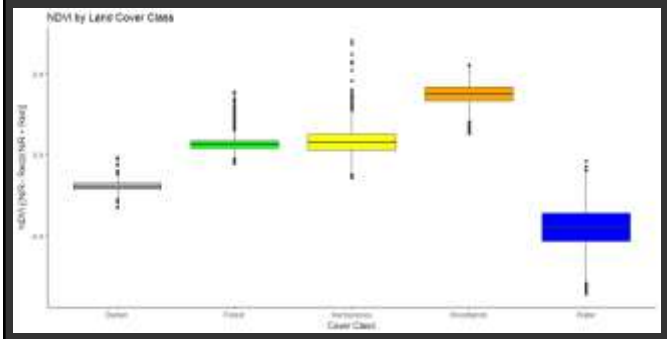
Boxplots can be used to compare numeric data between different groups. A nominal variable can be used to define different groups while a numeric variable defines what measure is being compared. On this slide, I am comparing stream runoff by month (runoff is a measure of the amount of water in the stream) and the gas mileage based on number of engine cylinders.

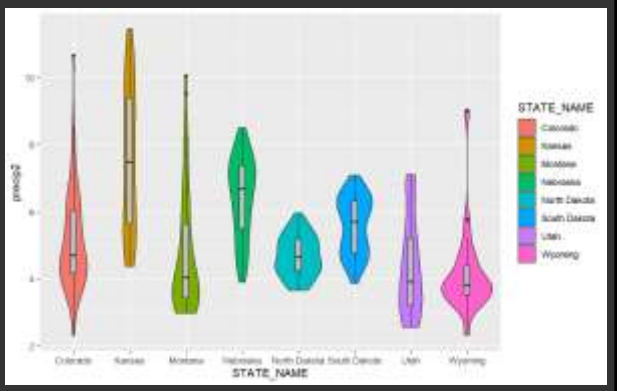
For the runoff data, generally the stream has less water running through it during the summer months and more during the spring. For the cars data, 4-cylinder vehicles tend to have better gas mileage than 6- or 8-cylinder vehicles.

I generally find box plots to be the best option for comparing numeric data between groups.

You may notice that there are some points in the graph space. These plots were created with ggplot2 in R, and this software will plot high values that are more than 1.5 IQR from the 3rd quartile or low points that are more than 1.5 IQR from the 1st quartile as outliers.

Comparing Groups with Box Plots and Violin Plots

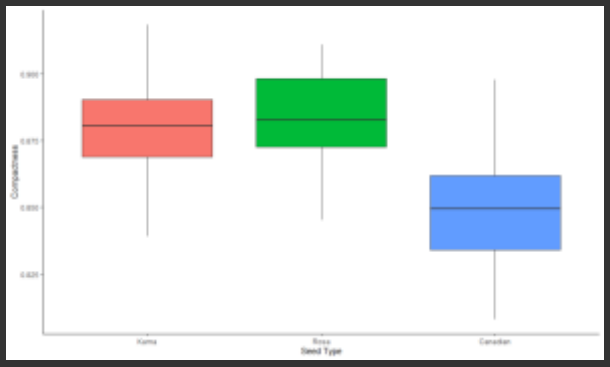




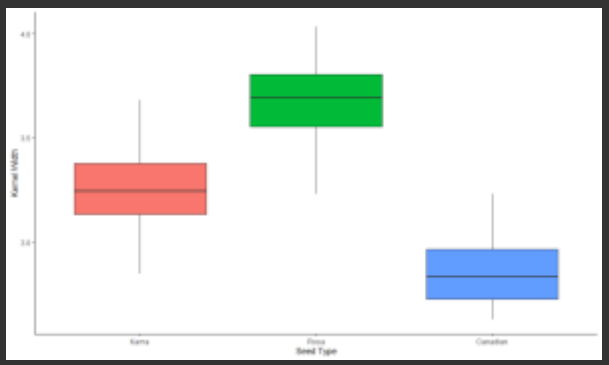
26

Note that a violin plot can be a substitute for or included with a box plot. A violin plot is a kernel density plot that has been turned on its side and mirrored. Since the x- and y-axis can be the same for a box plot and violin plot, they can be graphed in the same space. These plots can compliment each other and can provide a lot of collective information.

Box Plot Example: Seeds



- Compare the box plots on compactness and kernel width of the 3 seed types



27

These are examples of a box plots for the Seeds data comparing the compactness and kernel width for the three wheat varieties: Kama, Rosa, and Canadian. Generally, the Kama and Rosa varieties have similar compactness whereas the Canadian variety is more different form the other two. All three varieties tend to have unique kernel widths.

Summary of Key Points: Data and Graph Types

- Categorical data: summarize the number of features in each category or the proportion of percentage of features in each category
- Numeric data: summarize using different measures of central tendency and variability.
- Histograms and kernel density plots: useful for assessing the distribution of a single numeric variable
- Box plots and violin plots: useful for visualizing the distribution or a single numeric variable and can compare how the distribution of a numeric variable varies between groups.

Exploring Relationships

29

We will now shift our focus to discussing graphical means to explore relationships between variables, including a variety of bivariate or multivariate graphs.

Relationships: Categorical

- Do the categories of one variable correlate or depend on the categories of another variable?
- Summarize
  - Crosstabulation between two sets of categories
  - Contingency Tables
- Test
  - Chi-Square Test

		Variable 1		
		A	B	C
Variable 2	X	199	33	124
	Y	4	17	19
	Z	16	152	52

30

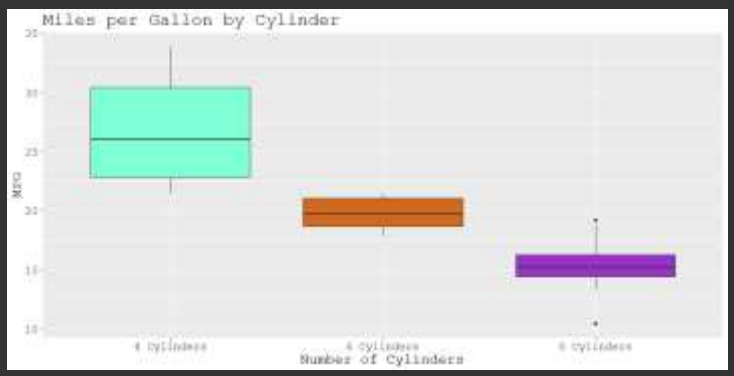
In order to explore the association between two categorial variables, it is common to use contingency tables. In a contingency table, one variable defines the rows while the other defines the columns. The cells hold the count of features that are included in each combination of levels. If there is no association or dependency between the variables, we should see no pattern in the counts. In the example table, it appears that features in Class A for Variable 1 are also likely to be in Class X for Variable 2. So, there seems to be an association here.

In short, contingency tables allow us to explore whether the categories of one variable correlate or depend on the categories of another variable.

From a contingency table, it is possible to perform a statistical test of dependence between two categorical variables using the Chi-Square Test. We will explore this test in a later module.

Relationships: Categorical and Numeric

- Do the values of a continuous variable vary between different groups?
- Summarize
  - Grouped Box Plot
  - Violin Plots
- Test
  - T-Test
  - Analysis of Variance (ANOVA)



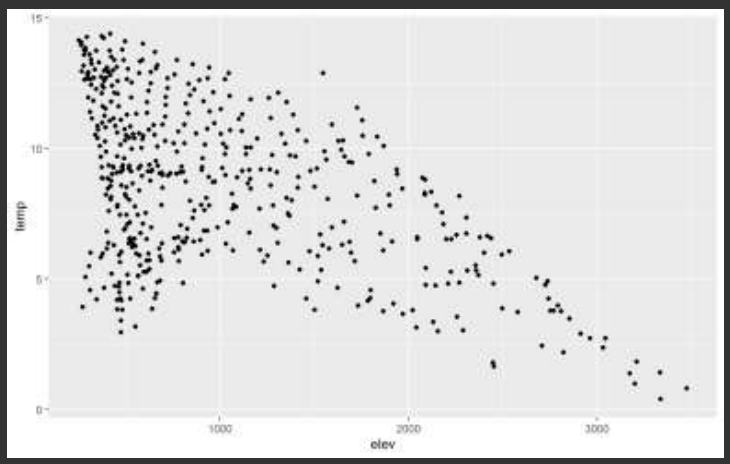
31

Box plots with groups separated and/or violin plots with groups separated offer a graphical means to explore whether the distribution of a numeric variable varies by group. In the example, it would visually appear that 4-, 6-, and 8-cylinder vehicles tend to have different fuel efficiencies. In short, this graphic allows us to explore whether values of a continuous or numeric variable are different between groups.

In a later module, we will explore statistical tests to compare the values of a continuous variable between groups. For example, the T-Test allows you to compare two groups while Analysis of Variance (ANOVA) allows you to compare more than two groups.

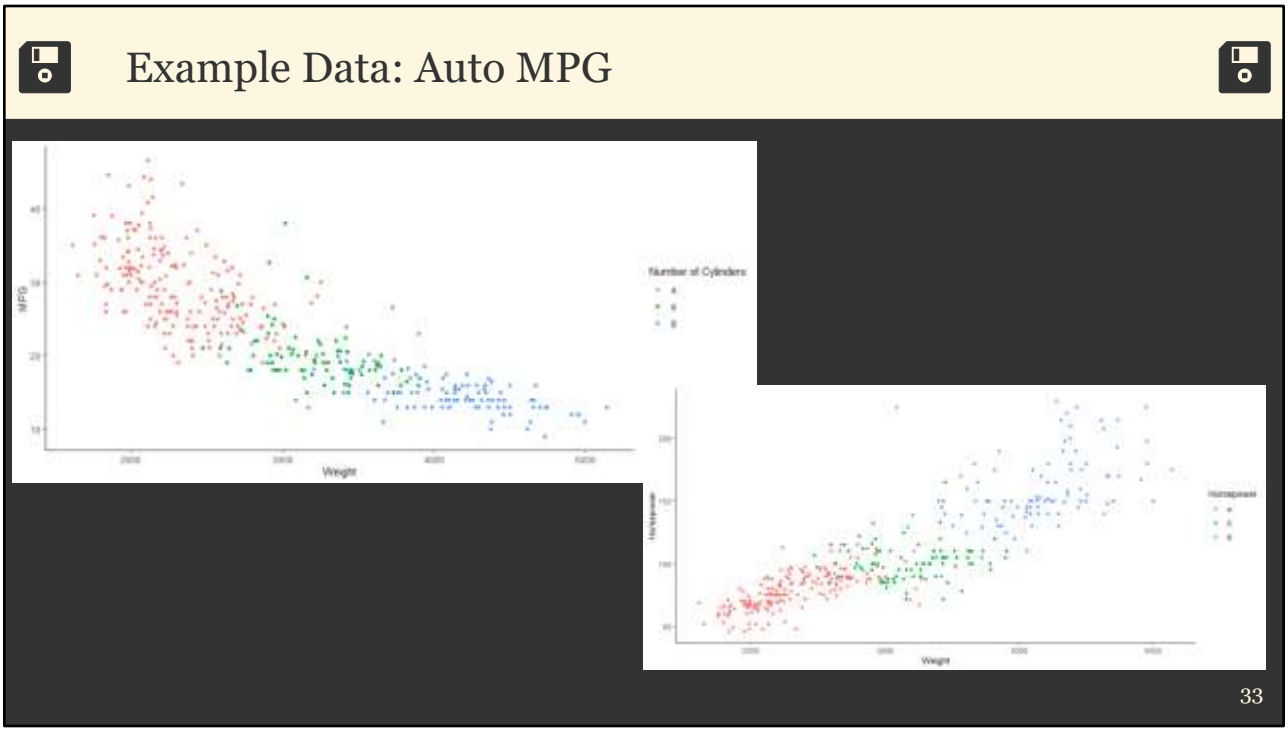
Relationships: Numeric

- Is there correlation or relationships between two continuous variables?
- Summarize
  - Scatter Plot
  - Scatter Plot Matrix
  - Line Graph
- Test
  - Correlation
  - Regression



32

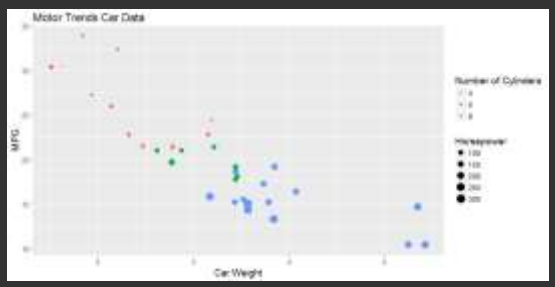
To visualize the relationship between two continuous variables, it is common to use a scatter plot. On this slide, I have provided a scatter plot for the High Plains data that has the elevation mapped to the x-axis position and the temperature mapped to the y-axis position. It appears as if there is a relationship here. Or, counties with higher mean elevation tend to have lower average mean annual temperatures. However, there is not a perfect relationship here, which suggests that other factors are of importance. For example, latitude may also play a role.

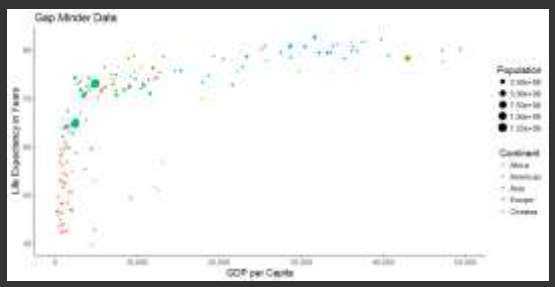


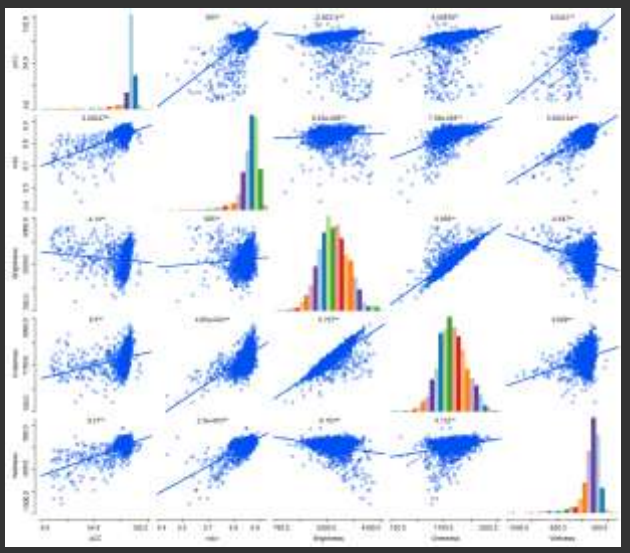
Here are some example scatter plots for the Auto MPG dataset. The plot on the left graphs weight on the x-axis and miles-per-gallon on the y-axis. Generally, this suggests an indirect, negative, or inverse relationship: heavier vehicles tend to have lower fuel efficiencies. In contrast, the graph on the right suggests that weight and horsepower tend to be directly or positively correlated.

Note also that I used unordered colors to show a nominal variable: 4-, 6-, vs. 8-cylinder vehicles. Generally, more cylinders tends to correlate with lower fuel efficiency, more horsepower, and higher weight.

Scatter Plot and Scatter Plot Matrix





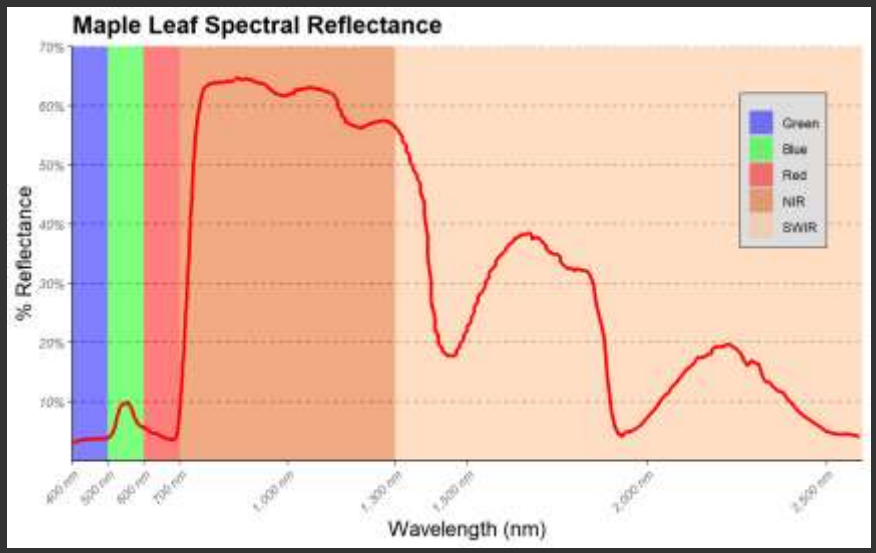


34

The scatter plots on the left side of the slide show additional aesthetic mappings. The top graph includes horsepower using the size of the symbol. The bottom graph is for the Gap Minder dataset. Here, gross domestic product per capita (GDP per capita) is mapped to the x-axis and life expectancy is mapped to the y-axis. Each dot represents a country. The size of the dot represents population, a continuous variable, and the unordered colors represent the continent on which the country occurs, a nominal variable. This is a bleak graph which is worth taking some time to digest.

Ignoring other aesthetic mappings, scatter plots only allow for the comparison of two numeric variables using position relative to the x- and y-axis. A scatterplot matrix allows for the comparison of a set of variables as a matrix of pairs. So, all variables of interest are graphed pairwise to allow for a visual assessment of correlation. The diagonal axis in the example provides histograms for each included variable. Scatter plot matrices are a common data exploration tool as they allow for correlations between multiple variables to be visually assessed.

Line Graph



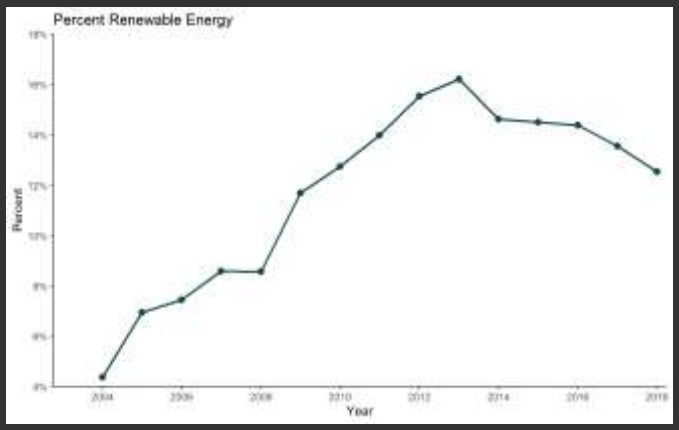
35

Line graphs connect data points using a line as opposed to mapping them as individual points. A line graph may be used as opposed to a scatter plot if there is a trend or series. The example provided on this slide is a spectral reflectance curve, which visualizes the amount of radiation that is reflected by an object at different wavelengths. This specific example is a spectral reflectance curve for a maple leaf. The x-axis is the wavelength, which varies from blue to shortwave infrared. The y-axis is percent reflectance. Larger values indicate that a higher percentage of that energy is reflected while lower values indicate less energy is reflected. For example, leaves tend to absorb blue and red wavelengths but reflect a little more green, which explains why they appear green. They also tend to reflect a lot of near infrared radiation, which does not impact how we perceive the leaf since we cannot see near infrared radiation.

A line graph makes more sense here because there is a trend or varying levels of percent reflectance with changes in wavelength. Plotting a line instead of individual points makes this trend clearer.

Time Series Graph

- X-Axis = Time
- Y-Axis = Continuous Variable



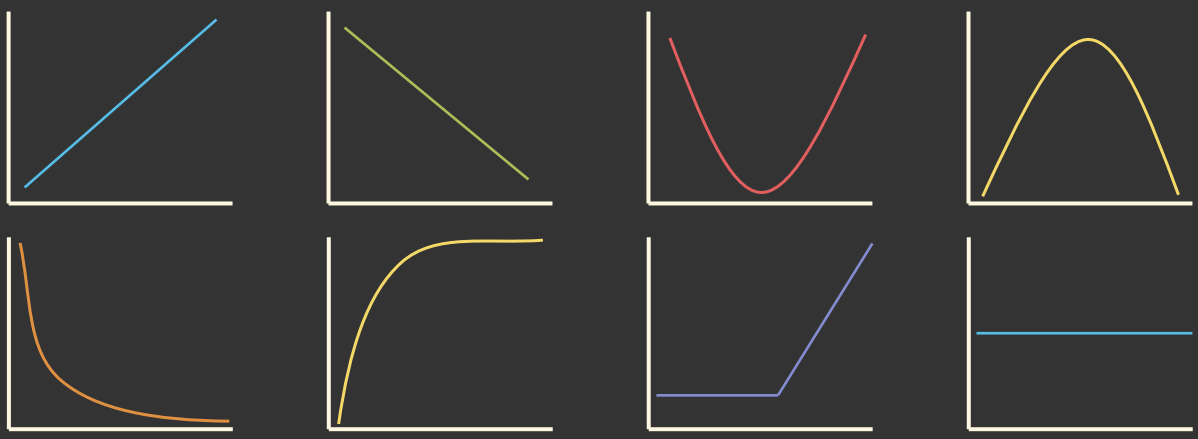
36

Some of the most common types of line graphs are time series graphs in which time maps to the x-axis and a variable that changes over time maps to the y-axis. In this specific example, we are mapping the percent of energy used by a country that is generated using renewable methods against time, which allows us to visualize potential trends with time.

There are many variables that vary with time and are often visualized using a time series graph. Examples include temperature, precipitation, population, stock market trends, etc.

Correlation

Correlation = Degree of association between two variables



37

The concept of correlation can be a bit confusing because its meaning can be vague. Specifically, there are different types of correlation between variables. Generally speaking, correlation relates to the degree of association between two variables. Or, if the value of one variable changes, then the value for the other variable is also likely to change.

The graphs on this slide represent different types of correlation. When a pattern or trend is observed when graphing two variables using a scatter plot then this suggests some degree of correlation exists. As these graphs demonstrate, there can be a wide variety of correlations or patterns. All of these graphs represent correlation except for the final graph at the bottom-right position. As the value on the x-axis varies, the values on the y-axis stays the same, indicating no correlation.

Pearson Correlation

Pearson = measure of linear correlation between two variables

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Assumptions

(1) normal distribution, (2) linearity, and (3) homoscedasticity

Equation from Wikipedia: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

38

One specific type of correlation is linear correlation. Linear correlation indicates that the correlation between two variables can be best modeled using a straight line with a positive slope for a positive relationship (one variable goes up, the other goes up) or a negative slope for a negative relationship (one variable goes up, the other goes down).

One measure of linear correlation is Pearson's Correlation Coefficient, or Pearson's r. The equation for this measure looks a bit complicated, but it is actually pretty simple. For each data point, the mean for Variable 1 is subtracted from the Variable 1 value. This is then multiplied by the difference between the Variable 2 mean and Variable 2 value for this sample. The result for each data point is then summed for all data points. This is the numerator of the equation. The denominator is effectively the product of the standard deviation of the two variables.

Pearson's Correlation Coefficients will range from -1 to 1. A value of -1 indicated a perfectly negative linear correlation whereas a value of +1 indicates a perfectly positive linear correlation. Values near zero indicate no correlation.

Note that this measure is designed to specifically assess whether a linear correlation exists. This is not a generic measure to assess any form of correlation. It also has some assumptions. We will discuss statistical

assumptions in detail in a later module.

Spearman Correlation

Spearman = measure of monotonic correlation based on ranks

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\operatorname{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}},$$

Assumption

(1) monotonic relationship

Equation from Wikipedia: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

39

Spearman's Rank Correlation, or Spearman's rho, offers another means to assess correlation between two variables. In this case, it is not assumed that the correlation is linear. However, it is assumed that it is monotonic. This means that an increasing or decreasing pattern in one direction is assessed. This measure would be misleading if the correlation or pattern is not monotonic, such as a relationship that increases to a value then decreases, such as a parabolic shape.

This measure relies on the ranks of the data points as opposed to the actual data values. Generally, two variables would be considered to be more correlated when the ranks of the two variables are similar for the same data point.

Since Spearman's Rank Correlation does not assume a normal distribution in the data, it is considered a nonparametric measure of monotonic correlation. We will discuss the concept of parametric vs. nonparametric tests in a later module.

Kendall Correlation

Kendall = measure of rank correlation based on sorted order

Concordant = Ordered the same way

Discordant = Not ordered the same way

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}.$$

Assumption

(1) monotonic relationship

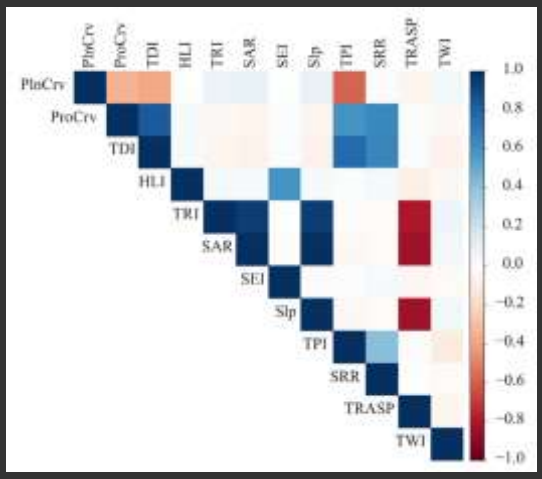
Equation from Wikipedia: https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient

40

Kendall's Tau offers another measure of nonparametric, monotonic correlation between two variables. This measure is based on the number of concordant (Variable 1 and Variable 2 have the same rank for a specific data point) and discordant (Variable 1 and Variable 2 do not have the same rank for a specific data point) pairs or data points.

Spearman's rho and Kendall's Tau tend to be related. However, they offer different means to assess for a monotonic relationship using ranks and nonparametric methods.

Correlation Matrix



	A	B	C
A			
B			
C			

Maxwell, A.E. and Shobe, C.M., 2022. Land-surface parameters for spatial predictive mapping and modeling. *Earth-Science Reviews*, p.103944.

41

When you have more than two numeric variables for which you want to compare correlation, this can be accomplished using a correlation matrix in which each pair of variables is compared separately. Since comparing Variable 1 to Variable 2 is the same as comparing Variable 2 to Variable 1, only half of the matrix needs to be populated.

On this slide, I have included an example graphic of a correlation matrix from one of my own studies in which we compared the correlation between different variables that are used to describe local topographic or terrain characteristics using Spearman's Rank Correlation. In this graph, darker red indicates a stronger negative correlation (one variable goes up, the other goes down) whereas darker blue indicates a positive correlation (one goes up, the other goes up). The bottom diagonal cells all show a correlation of 1 because the variable is being compared with itself. By definition a variable is perfectly correlated with itself.

Example Data: Seeds

Pearson

	area	perm	compactness	lkernel	wkernel	asymn	lgroove
area	1.0000000	0.9943409	0.6082884	0.9499854	0.9707706	-0.22957233	0.86369275
perm	0.9943409	1.0000000	0.5292436	0.9724223	0.9448294	-0.21734037	0.89078390
compactness	0.6082884	0.5292436	1.0000000	0.3679151	0.7616345	-0.33147087	0.22682482
lkernel	0.9499854	0.9724223	0.3679151	1.0000000	0.8604149	-0.17156243	0.93280609
wkernel	0.9707706	0.9448294	0.7616345	0.8604149	1.0000000	-0.25803655	0.74913147
asymn	-0.2295723	-0.2173404	-0.3314709	-0.1715624	-0.2580365	1.00000000	-0.01107902
lgroove	0.8636927	0.8907839	0.2268248	0.9328061	0.7491315	-0.01107902	1.00000000

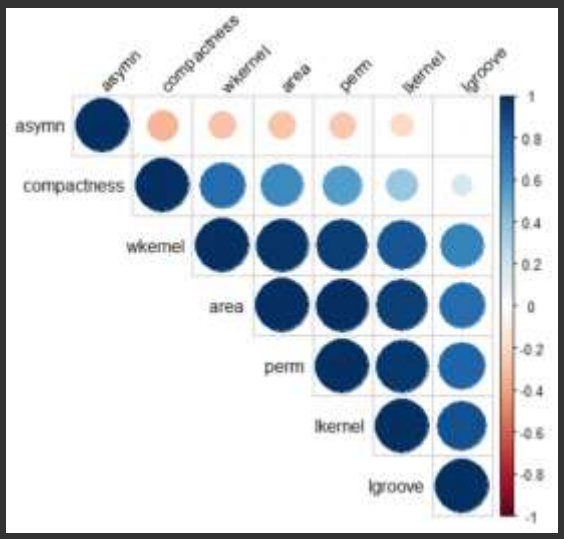
Spearman

	area	perm	compactness	lkernel	wkernel	asymn	lgroove
area	1.0000000	0.9913799	0.6384959	0.9309151	0.9748895	-0.27807829	0.76181635
perm	0.9913799	1.0000000	0.5532552	0.9627474	0.9456918	-0.26067246	0.80285817
compactness	0.6384959	0.5532552	1.0000000	0.3780784	0.7627271	-0.33884242	0.16137935
lkernel	0.9309151	0.9627474	0.3780784	1.0000000	0.8501443	-0.19924546	0.87541557
wkernel	0.9748895	0.9456918	0.7627271	0.8501443	1.0000000	-0.28209605	0.66973936
asymn	-0.2780783	-0.2606725	-0.3388424	-0.1992455	-0.2820960	1.00000000	0.02050324
lgroove	0.7618164	0.8028582	0.1613794	0.8754156	0.6697394	0.02050324	1.00000000

The outputs presented on this page provide correlations using the Pearson and Spearman methods for the numeric variables included in the Seeds data. Note that the entire matrix has been populated, so measures repeat on the opposite side of the diagonal.

Again, larger negative values indicate a stronger negative correlation. For example, as kernel compactness increases, kernel asymmetry tends to decrease. In contrast Kernel length and kernel width tend to be highly positively correlated.

Example Data: Seeds



43

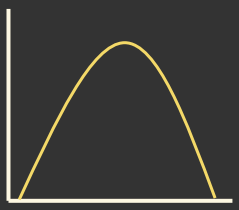
This slide provides a visualization of the Spearman's Rank Correlation measures for the Seed data. Larger circles indicate a stronger positive or negative monotonic correlation based on ranks. Darker red indicates a stronger negative correlation, and darker blue indicates a stronger positive correlation.

Generally, this plot suggests many of the variables are positively correlated. However, Asymmetry tends to be negatively correlated with the other variables, other than groove length.

Measuring Nonlinear or Non-Monotonic Correlation

Wang, Q., Shen, Y. and Zhang, J.Q., 2005. A nonlinear correlation measure for multivariable data set. *Physica D: Nonlinear Phenomena*, *200*(3-4), pp.287-295.

- Available methods are more limited
- Consider fitting specific functions to data and assessing fit
  - Sin
  - Cos
  - Polynomial
  - Exponential
  - Etc.



Available online at www.sciencedirect.com

SCIENCE DIRECT

ELSEVIER

PHYSICA D

A nonlinear correlation measure for multivariable data set

Qiang Wang[a,*], Yi Shen[a], Jian Qiu Zhang[b]

44

As described above, Pearson's Correlation specifically assesses linear correlation whereas Spearman's and Kendall's Correlation assess monotonic correlation based on ranks. In other words, all of these measures assess correlation that is represented as decreasing or increasing in one direction. However, there are other forms of correlation, such as an oscillating pattern.

The small graph provided on this page represents a strong relationship between two variables that is not monotonic, or increasing in one direction. Instead, as the variable mapped to the x-axis increases, the variable mapped to the y-axis also increases until the x-variable reaches a specific value. At that point, as the variable mapped to the x-axis increases, the variable mapped to the y-axis decreases. So, this pattern would represent a correlation, but not a monotonic correlation. How are such relationships assessed?

Unfortunately, measures of non-linear or non-monotonic relationships are more limited and there is not a general consensus as to the best or most appropriate method. The paper highlighted on this slide presents one proposed method. Another option is to fit a function to the data that matches the relationship for which you are interested in assessing correlation. For example, the graph shown on this slide could be represented using a polynomial equation. Oscillating patterns might be modeled using a sine or cosine curve. You can then assess how well this mathematical function approximates the

relationship.

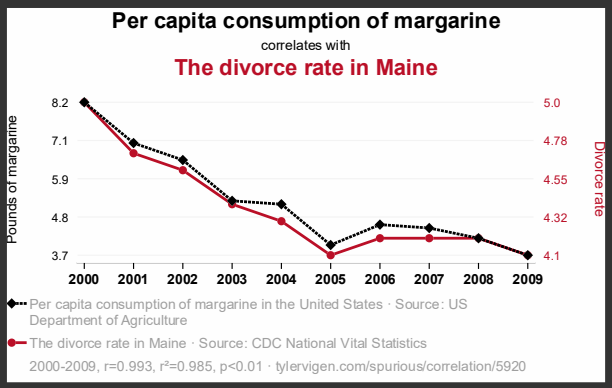
Such methods are outside the scope of this course. However, I wanted to make it clear what the Pearson, Spearman, and Kendall measures are designed to assess regarding certain types of correlation: linear in the case of Pearson and monotonic in the case of Spearman and Kendall. Other types of correlation exist that these methods are not designed to measure. In short, it is important to understand what is meant by correlation in the context of interest and in the context of the metrics used to assess it.

Correlation vs. Causation

Relationships do not always imply that one variable causes response in the other

Correlations are not always meaningful.

Correlation does not imply causation!



https://www.tylervigen.com/spurious-correlations

45

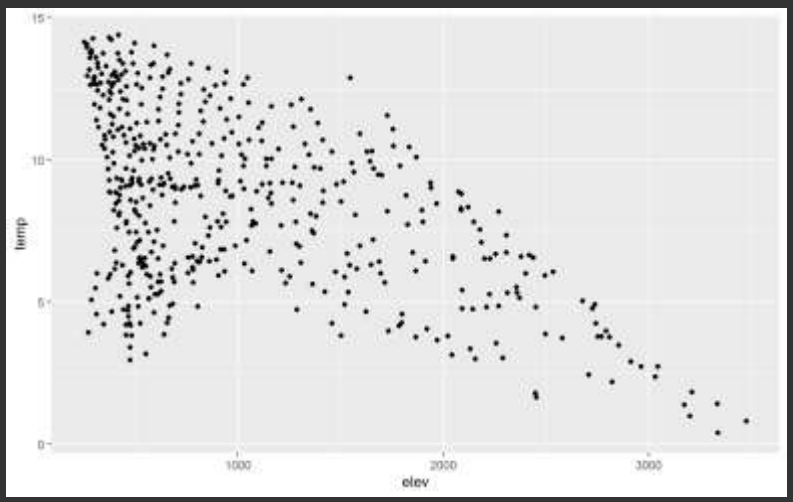
A common mistake is to conclude that a correlation between two variables indicates one variable causes a response in the other variable. Relationships between two variables do not always imply that changes in one variable causes changes in the other. Also, correlations are not always meaningful. Correlation does not imply causation. The website linked on this slide provides some interesting examples of spurious correlations. Here, the graphics indicate that there is a correlation or relationship between the variables. However, this is purely coincidental and meaningless.

When you do observe a correlation that is not spurious or meaningless, this does not necessarily imply causation. For example, you may find that there is a positive correlation between skin cancer rates in a geographic area and the number of oranges consumed by individuals living in that area. However, the likely cause of this correlation is that both skin cancer and orange consumption are correlated with sun exposure since sun exposure can lead to skin cancer and oranges tend to grow in warm, sunny climates.

So, be careful when making assumptions of causation when you observe a correlation between variables.

Dependent and Independent Variables

- Dependent Variable = Variable that depends on or responds to the independent variable
- Independent Variable = Variable that impacts or causes a response in the dependent variable



46

When you are creating a graph, how do you decide which variable to map to the x-axis and which to map to the y-axis? Generally, the independent variable is mapped to the x-axis and the dependent variable is mapped to the y-axis. The dependent variable is the variable that depends on or responds to the independent variable whereas the independent variable impacts or causes a response in the dependent variable.

In the example graph, it makes since to map elevation to the x-axis since temperature is likely partially dependent on elevation as opposed to elevation being partially dependent on temperature.

Sometimes it is not clear which variable is the dependent variable and which is the independent variable. For example, you may simply want to visualize the correlation and not identify a dependent and independent variable. In such situations, which variable is mapped to the x- vs. y-axis is somewhat arbitrary.

When time is one of the variables being explored, it is commonly mapped to the x-axis to generate a time series graph, and it is assumed that the other variable is dependent on or changes with time. However, there are situations where time should not be mapped to the x-axis.

In later modules in the class, we will discuss linear regression and predictive

modeling. Generally, these techniques are used to predict a dependent, or y, variable using one or multiple independent, or x, variables. In these cases, the dependent variable should be mapped to the y-axis.

Summary of Key Points: Exploring Relationships

- The dependency or relationship between two categorical variable can be summarized in a contingency table.
- Box plots can be used to visualize the difference in the distribution of a numeric variable between groups.
- Correlation between continuous variable can be visualized using scatter plots.
- For more than two variables, additional aesthetic mappings can be used, such as size to show another numeric variable or color to how a numeric or nominal variable.
- The Pearson's Correlation is a measure of linear correlation. Spearman and Kendall correlation are a measure of monotonic correlation based on ranks.
- Assessing relationships that are not linear or not monotonic are generally more complex
- Not all correlations are meaningful. Correlation does not imply causation!

This is the end of this lecture module.

Please return to the West Virginia View Webpage for additional content.

Thanks! Hope you found this useful.