Graphical Grammar

Brian Vanover

Xuan Yang

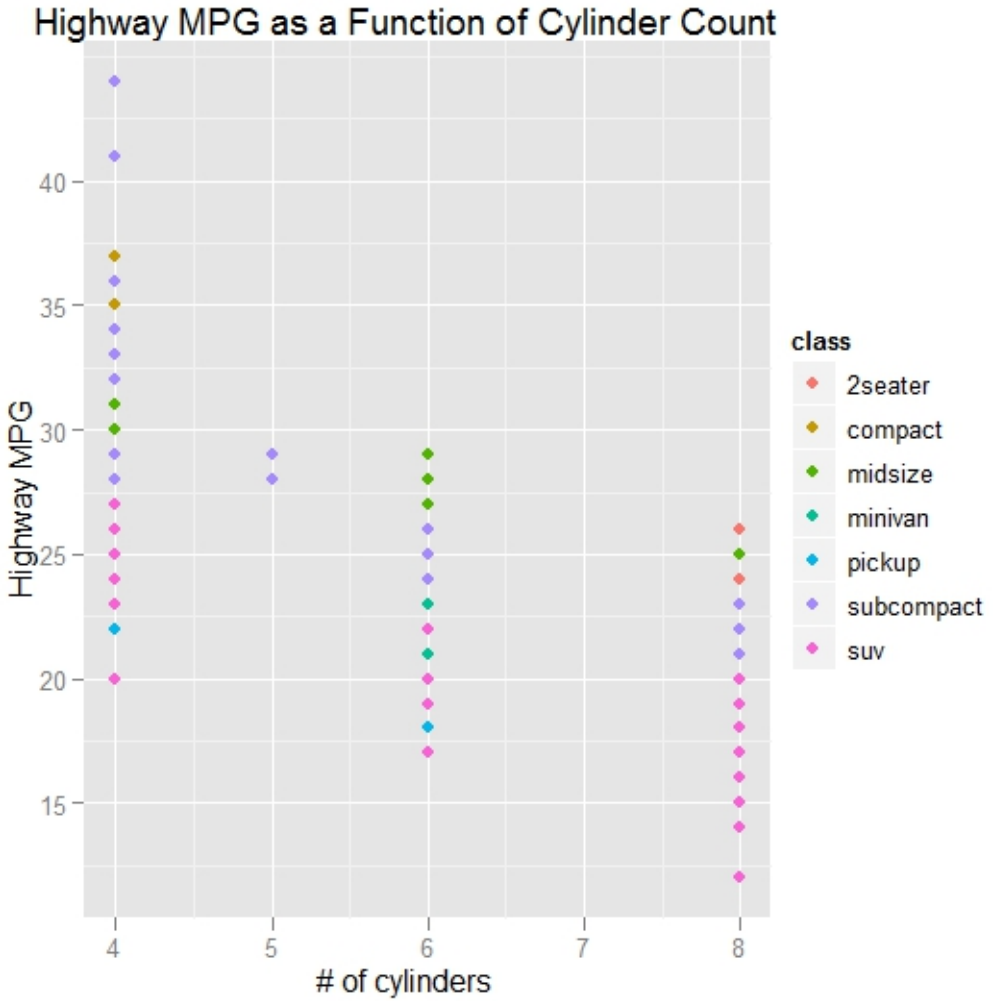
01/19/2011

Build a plot

- Many different types of plots.
- Convert *data* units to *physical* units
- Scale and Statistically transform the data
- Combine graphical objects from 3 sources
  1. Data
  2. Scales and Coordinate System
  3. Plot Annotations (Title, background)

See Example

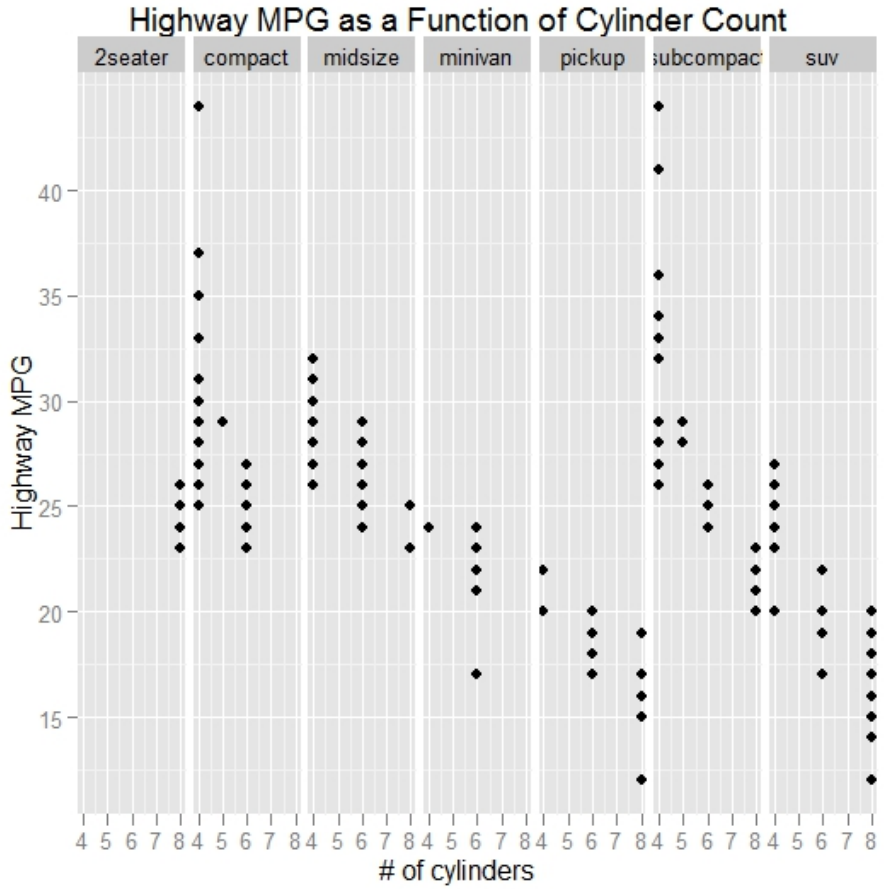
Can you think of other ways to represent this information graphically ?



Faceting

- “Produces small multiples showing different subsets of the data.”
- Scaling occurs in three parts
  1. Transforming
     - Occurs before stat transformation
     - Only necessary for non-linear scales
  2. Training
     - Combines ranges of datasets to get complete range
     - Locally applied scales → Meaningless comparisons
  3. Mapping
     - Map data values to aesthetic values
     - Easier to map within each facet as opposed to splitting final

Faceting by Class.
Discuss the intuitive process used to build this plot.

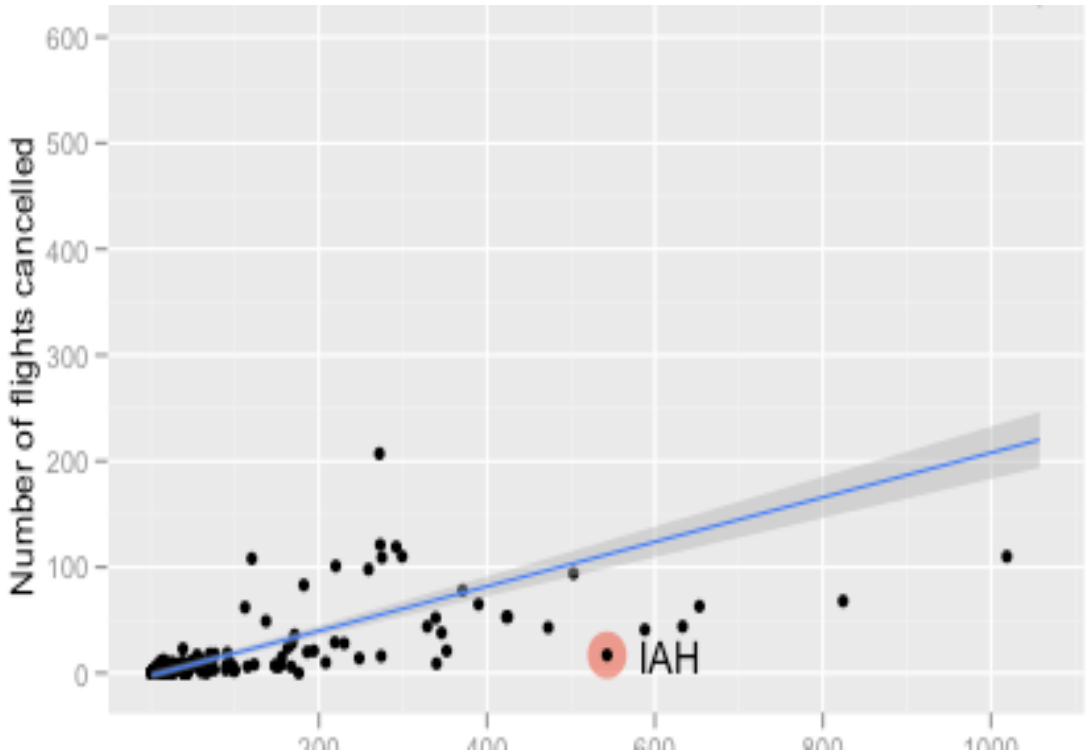


Components of Layered Grammar

- Default Dataset, Set of mappings from variables to aesthetic
- One or more layers each having
  - One geometric object, statistical transformation, position adjustment, and dataset/set of aesthetic mappings
- One scale for each aesthetic mapping used
- A coordinate system
- The facet specification

Benefits/Characteristics

- Components are independent
- Layer component determines physical representation of data
- Grammar makes iterative plot updates easier
  - Suggests ways plots *can* be changed
  - Promotes creation of new/customized graphics



An Example of Layers and Their Defaults

```
ggplot(feb13, aes(ntot, ncancel))
  + geom_point(data = subset(feb13, origin == "IAH"), size = 7, colour = alpha("red", 0.5))
  + geom_point()
  + geom_text(data = subset(feb13, origin == "IAH"), aes(label = origin), hjust = -.5)
  +  geom_smooth(method = "lm", se = T) + labs(y = "Number of flights cancelled", x = "Total number of flights")
```

Component Characteristics

- Data & Mapping
  - Can construct graph applicable to multiple dataset
  - Specify which variables are mapped to which aesthetics
- Statistical Transformation
  - Transforms data, typically by summarization
  - Must be location-scale invariant
- Geometric Object
  - Control type of plot created
  - Classified by dimensionality
  - Every geom has default statistic vice versa
  - Can only display certain aesthetics

Can you guess the accompanying default geoms for these given statistics?

1. Bin
2. Boxplot
3. Identity
4. Contour
5. Smooth

Characteristics Cont.

- Position Adjustment
  - Tweak position of geom objects that obscure others
- Scales
  - Controls mapping from data to aesthetics
  - Need one scale for each aesthetic used in a layer
  - Consists of a function, its inverse, and set of parameters
- Coordinate System
  - Maps position of objects onto plane of plot
  - Affect all position variables simultaneously and change appearance of geometric objects
  - Controls how axes and gridlines are shown
- Faceting

Hierarchy of Defaults

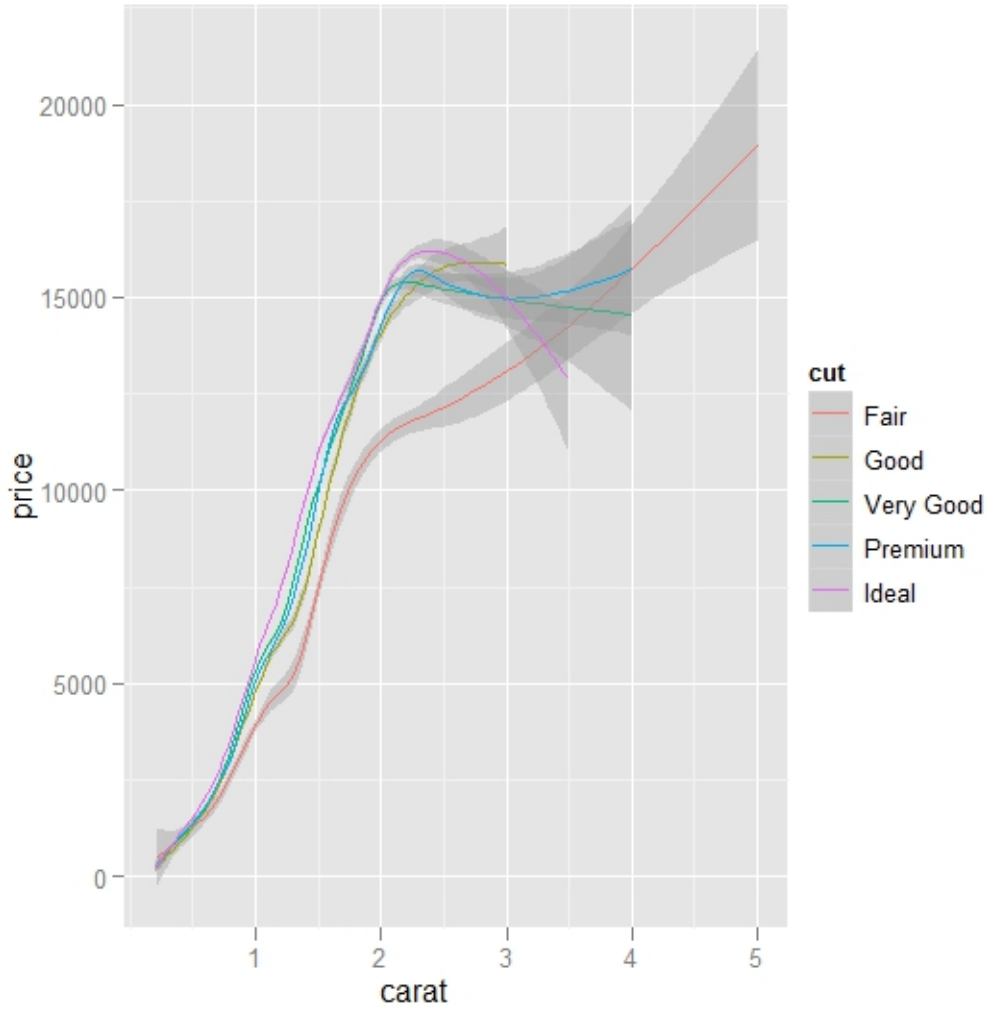
- Describing every component every time is a poor use of time
- Defaults simplify work of plotting
- Intelligent default
  - Need only specify one geom or stat
  - Cartesian coordinate system
  - Scales defaulted according to type of variable and aesthetic
  - Position-based mapping
- Qplot
  - Assumes multiple layers use same data/aesthetic
  - Defaults to scatterplot
  - Mimics syntax of R *plot* function

Intelligent Default and Qplot

We can construct the same graphic with the two following codes:

```
qplot(carat, price, data = diamonds, colour = cut, geom = "smooth")
```

```
plot3 <- ggplot(data = diamonds,   mapping = aes(x = carat, y = price, colour = cut)) +
 layer(data = diamonds, mapping = aes(x = carat, y = price, colour = cut),
 geom = "smooth", position = "identity", stat = "smooth") +
scale_x_continuous() +
scale_y_continuous +
coord_cartesian()
```

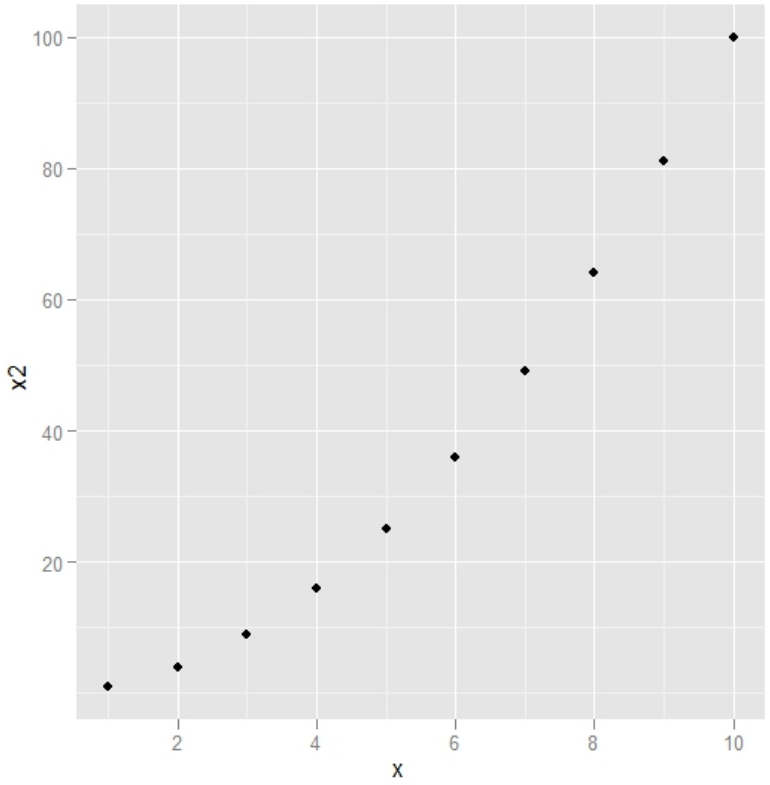


Implications of Layered Grammar

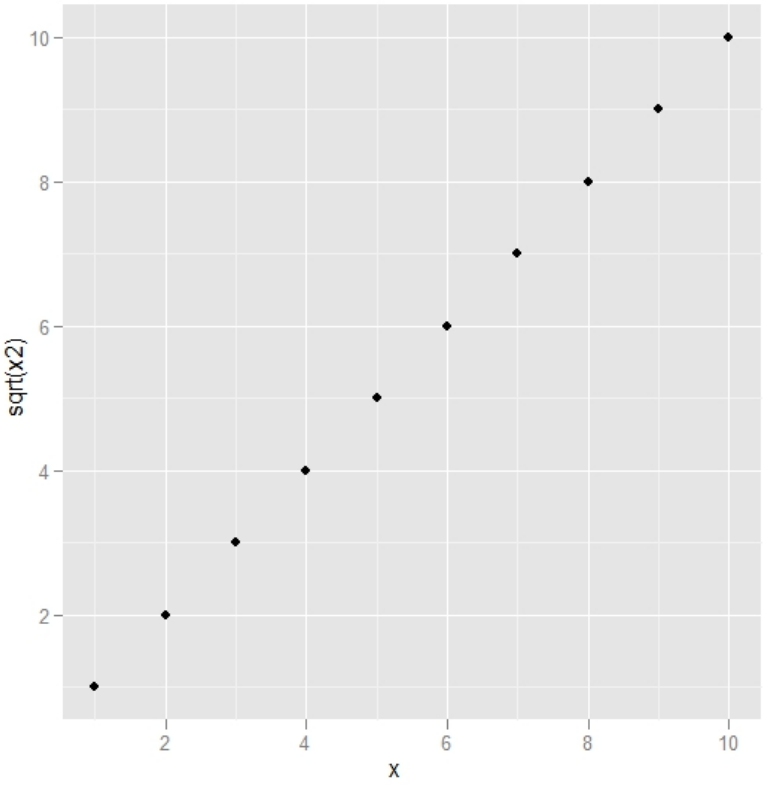
- Histograms
  - Default binwidth, and the choice of bins
  - Y-position not present in original data ..count..
- Polar Coordinates
- Transformations
  1. Data
  2. Scales
  3. Coordinate System

Transforming the Data

Data

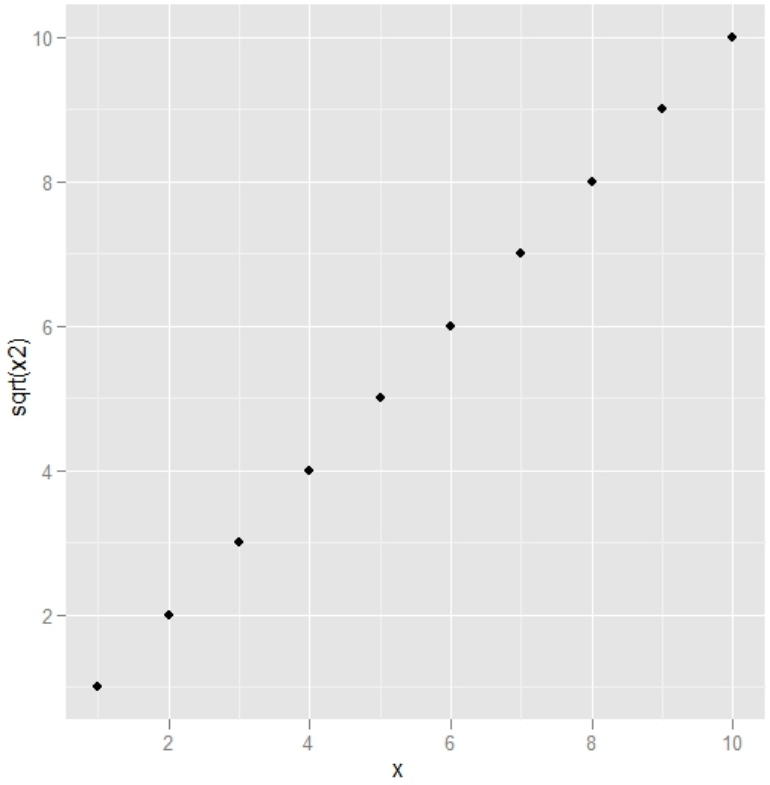


Transformed Data

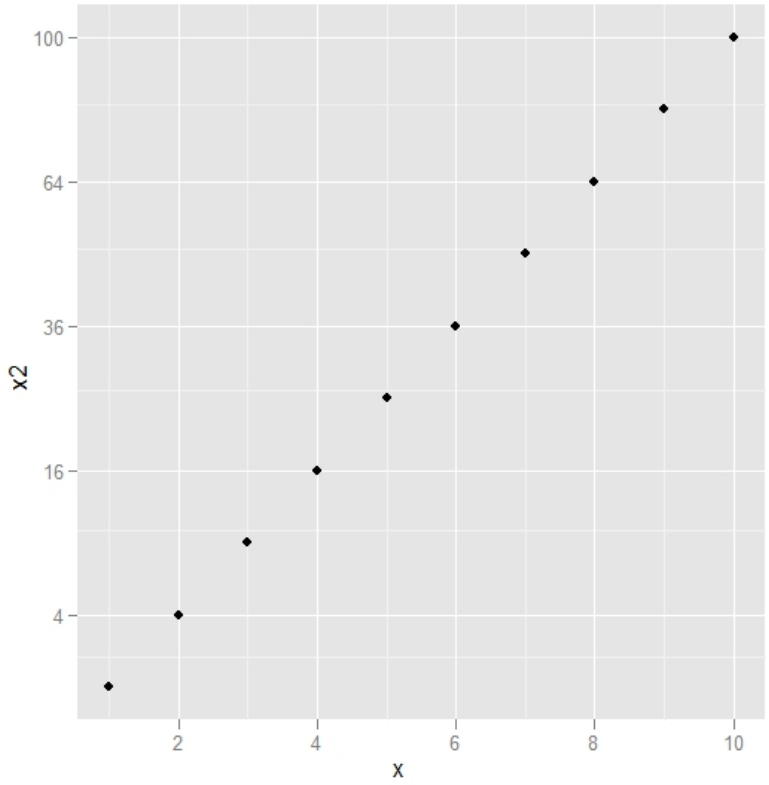


Transforming the Scales

Transformed Data

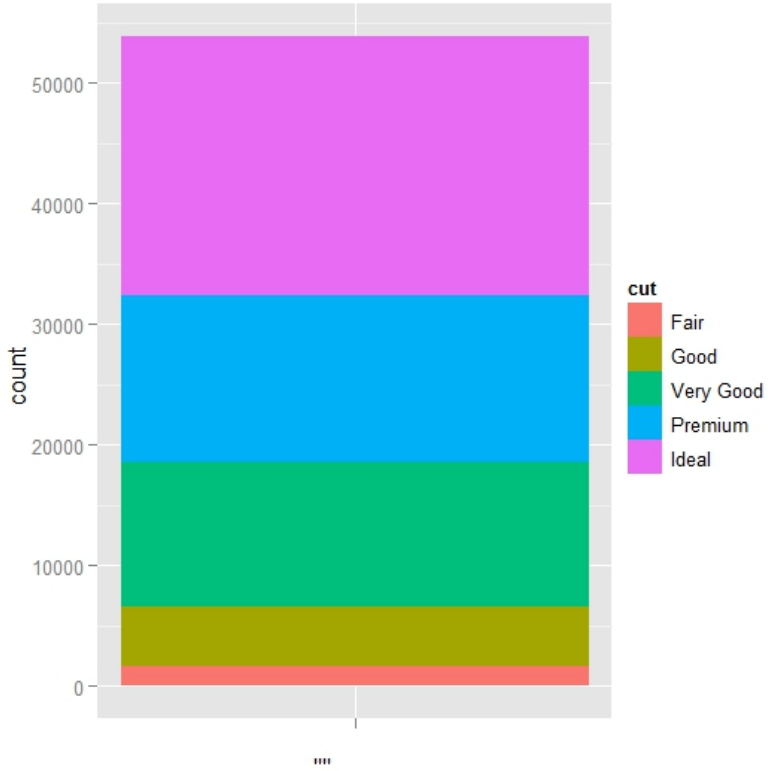


Transformed Scales

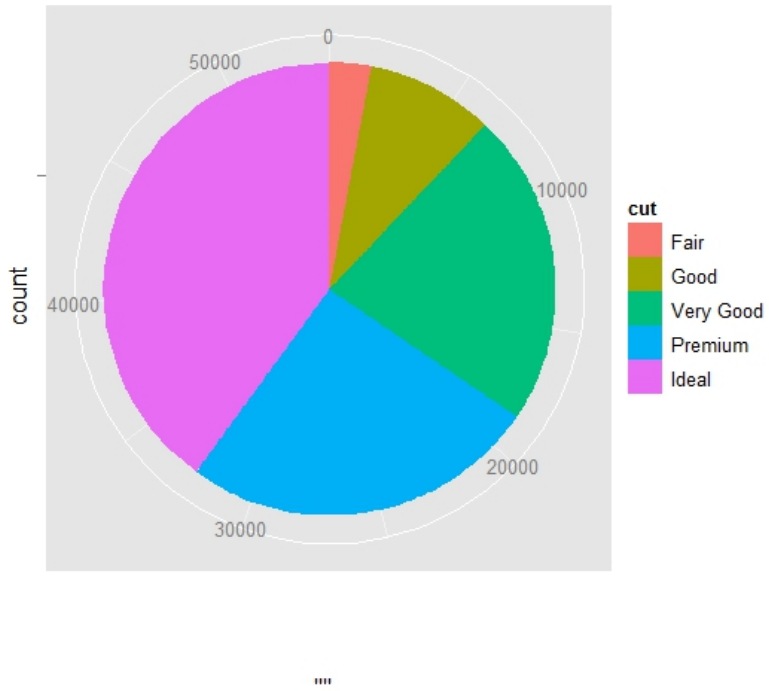


Transforming the Coordinate System

Cartesian Coordinates



Polar Coordinates



Common Mistakes; Possible Solutions

- Too many variables
  - Hard to see relationships between more than three variables, two position and one other
  - Warn the user and suggest alternatives such as faceting
- Overplotting
  - Prompts incorrect conclusions about distribution
  - Supplement plot with contours or color by density
- Alphabetical Ordering
  - Categorical variables often ordered alphabetically
  - Ordering by some property of data more useful
- Polar Coordinates
  - Humans better at judging length than angle or area
  - Difficult to judge an angle for objects with small radius

What are some other common mistakes?

Conclusions

- Aim is to “bring together in a coherent way things that previously appeared unrelated and which also will provide a basis for dealing systematically with new situations.”
- Layered grammar allows for more interchangeability, faster duplication, easier exploration of new graphics
- Grammar not so strong in area plots
  - Development of subgrammar
- Interactive plots
  - Binwidth slider
  - Speed
- Grammar is powerful and useful, but more specification of subgrammars and measures to ensure good graphics are needed

The Good and Bad of Graphics