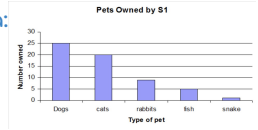
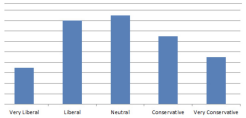
Categorical Data Visualization

Types of Categorical Data:



- **Nominal**: no fixed category order
  - Must be sorted by decreasing frequency
  - **Examples**: Color, Artists, Fruits



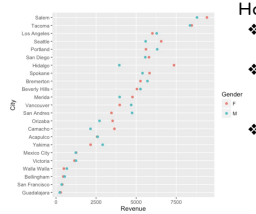
- **Ordinal**: fixed category order
  - Must be sorted in logical order
  - **Examples**: Educational level, satisfaction rating, feelings

- Discrete: small # of possibilities
  - **Examples**: # of children
- Data can be received in three different types of formats:
  - Cases, Counts, Contingency table

Cleveland dot plot

When to use:

- Large amount of categories
- Alternative to bar chart
- **Benefits**: Compact, Can plot multiple graphs on the same line



How to plot:

- R has a built in function called dotchart()
- For ggplot, you can use geom_point() and build your own cleveland dot plot from scratch
- Must change scales = "free_y" and space = "free_y" when faceting in order to create more understandable charts

Fluctuation Diagram

When to use:

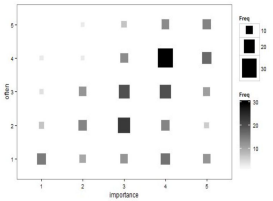
- Certain combinations of categories are rare and should not occur at all
- This graph is plotted to find such outliers
- **Benefit**: Can easily identify rare outliers

How to plot:

- Can use R function fluctile()

How to read the diagram:

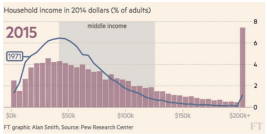
- Size of a point is mapped to the number of observations that fall within that bin



General Tips

Topcoding:

- When there is not enough data in top category
- Warning: if there is a large amount of data in top category, does not show details and wont understand spread



Faceting:

- View **multiple categories** in same graph
- facet_wrap( ) and facet_grid( )

NAs

- Create separate column for NA
- Make sure NA column is not overly prominent (put near bottom of bar chart) by making it a real factor level with fct_explicit_na ( )

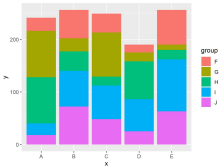
Recoding

- Use factor data instead of character data to more easily reorganize order of variables
- fct_relevel ( ): ordinal for manual reorder
- fct_reorder ( ): binned, nominal
- fct_infreq ( ): unbinned, nominal

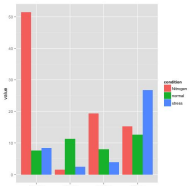
Bar chart

Stacked Bar Chart

- **Benefit**: Useful for emphasizing total of group
- **Detriment**: Difficult to tell count of top bar
- Can create **relative frequency stacked bar chart** for proportions

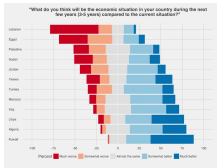


Grouped Bar Chart



- **Benefit**: Item-wise comparison within a group
- **Detriment**: Difficult to tell total count
- Can facet instead of using legend colors

Diverging Bar Chart



- Useful for **likert data**
- Neutral category can be in middle or off to the side

Mosaic plots

Filled rectangular plot with consistent number of rows and columns, where each small rectangle represents a unique combination of levels of factors of the variables displayed

- **Group sizes** determine size of each block
- Equivalent to a relative frequency stacked bar chart with unequal width
- Useful for understanding **associations**
- **Null hypothesis**: No relationship between variables (straight lines run through plot)

Best Practices:

- Vertical cut: independent variable
- Horizontal cut: dependent variable
- Most important level is closest to x-axis and darkest color

Variety:

- **Mosaic pairs plot**.
- **Mosaic spine plot**: straight, parallel cuts in one dimension and only one variable cutting in other direction