Data Visualization Categorical Data Cheat sheet



Definitions

What is categorical data?

A categorical variable is a variable that can take on one of a limited, and usually fixed number of possible values. There are three types of data.

Nominal variables have no fixed category order. You should sort the graph from highest to lowest count/value/frequency.

Ordinal variables have natural, ordered categories and the distances between the categories is not known. You should preserve the ordinal level of the variable when plotting a graph.

Real discrete numbers where the distance between categories are known. For example, child births per family.

One Variable

Small to moderate # of categories

Bar Charts

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally.

a <- ggplot(mpg) Nominal



a + geom_bar()



a + geom_bar(aes(fct_infreq
(manufaturer)))



a + geom_bar(aes(fct_rev(fct_infreq (manufacturer))))

Large # of categories

Cleveland Dot Plot

A Cleveland dot plots are a great alternative to a simple bar chart, particularly if you have more than a few items. It doesn't take much for a bar chart to look cluttered. In the same amount of space, many more values can be included in a dot plot, and it's easier to read as well.

b <- ggplot(USSeatbelt, year %in% 1997)



b + geom_point(aes(x=fatalities, fct_ reorder(fatalities, state)))

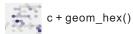
Two Variables+

Frequency Heatmap (2)

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors.

c <- ggplot(SpeedSki, aes(Year, Speed))



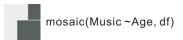


Association & Proportion

Mosaic Plot (2+)

A mosaic plot is a graphical method for visualizing data from two (or more) qualitative variables. It gives an overview of the data and makes it possible to recognize relationships between different variables.







Fluctuation Diagram (2)

The fluctuation plot is a variant of the mosaic plot. A grid is created so that is each combination of categories has a fixed position on the grid. Rectangle is drawn in a foreground color with a height proportional to the count for that particular combination of categories.



More info on how to plot a fluctuation plot is available online

Likert Data

Stacked Bar Chart

A likert data is an ordinal data that are usually responses from survey or questionnaire, used to measure attitudes or opinions. Five to seven items are usually used in the scale. For example:

- ·Strongly agree
- ·Agree
- ·Neutral
- ·Disagree
- ·Strongly disagree



HH::likert(GROUP~, yourdata, positive.rder)

