Comparing Proportions

Very often we want to compare the relative proportions of several quantities that may have ben measured at different times, places or contexts.

We might want to visualise how proportions of two quantities vary according to the values of another variable.

As a running example, I will visualise a dataset downloaded from the Central statistics office website. It shows the number of people in Ireland in 2016 that have self-reported in the 2016 census having a Masters of science or PhD degree or higher.The data is segregated by gender and age group.

I want to produce a visualisation that shows the proportional difference by gender of people holding higher degrees within the various age groups.

## Stacked Bars in Series

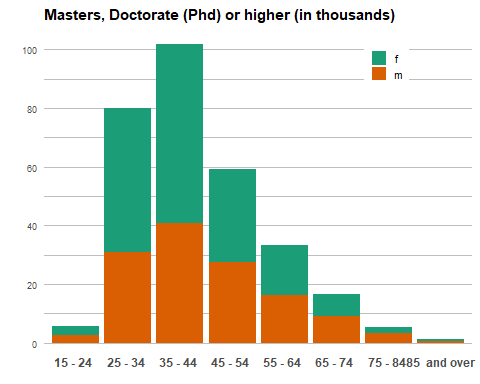
We can easily create a series of stacked bars representing the proportion of female to male degree holders per age group. You can achieve the simple Stacked Bar Graph simply by omitting the dodge attribute from geom\_bar

The objective of a plot like this is not only to show the proportions per age-group but to be able to compare these proportions across the age groups.

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.4

ire\_education <- read.csv("ire-2016-Masters, Doctorate (Phd) or higher.csv")  
  
ggplot(ire\_education, aes(x = age, y=number, fill=gender)) +  
 geom\_bar( stat="identity") +  
   
 scale\_y\_continuous(name = "", breaks = seq(0,100000, by=20000), labels = seq(0,100, by=20)) +   
   
 #scale\_fill\_brewer(palette = "Set1", name = "Uploads") +  
   
 scale\_fill\_manual(values = c( "#1b9e77", "#d95f02"),  
 name = NULL) +  
   
 ggtitle("Masters, Doctorate (Phd) or higher (in thousands)") +  
   
 theme\_classic() +  
   
 theme(  
 axis.line.x = element\_blank(),  
 axis.ticks.x = element\_blank(),  
 axis.ticks.y = element\_blank(),  
 axis.line.y= element\_blank(),  
 axis.title.x=element\_blank(),   
 axis.text.x = element\_text( vjust = 5, size=9, face="bold"),  
 axis.text.y = element\_text( size=7),  
 axis.title.y = element\_text(size=9, face="bold"),  
 legend.text = element\_text(size=8),  
 legend.title = element\_blank(),  
 legend.position = c(0.8,0.9),  
 legend.key.size = unit(0.8,"line"),  
 panel.grid.major.y = element\_line(size = 0.1, linetype = 'solid', colour = "grey"),  
   
 panel.grid.minor.y = element\_line(size = 0.1, linetype = 'solid', colour = "grey"),  
 plot.title=element\_text( hjust=0.00, face='bold', size=11)  
   
 )

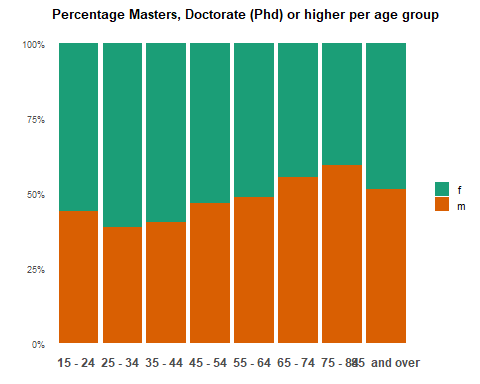
 ### 100% or Proportional Stacked Bar Graph

It may be useful to show the proportion out of 100% per age group. This is known as a proportional stacked bar graph or a 100% stacked bar graph.

To do this, you can use geom\_col instead of geom\_bar and set the position attribute to fill. With position = fill, the y values are scaled from 0 to 1.

To print the labels as percentages set labels = scales:percent in scale\_y\_continuous. You should reference the scales library first though

library(scales)  
  
  
ggplot(ire\_education, aes(x = age, y=number, fill=gender)) +  
 geom\_col(position="fill") +  
   
 scale\_y\_continuous(name = "", labels = scales::percent) +   
   
 #scale\_fill\_brewer(palette = "Set1", name = "Uploads") +  
   
 scale\_fill\_manual(values = c( "#1b9e77", "#d95f02"),  
 name = NULL) +  
   
 ggtitle("Percentage Masters, Doctorate (Phd) or higher per age group") +  
   
 theme\_classic() +  
   
 theme(  
 axis.line.x = element\_blank(),  
 axis.ticks.x = element\_blank(),  
 axis.ticks.y = element\_blank(),  
 axis.line.y= element\_blank(),  
 axis.title.x=element\_blank(),   
 axis.text.x = element\_text( vjust = 5, size=9, face="bold"),  
 axis.text.y = element\_text( size=7),  
 axis.title.y = element\_text(size=9, face="bold"),  
 legend.text = element\_text(size=8),  
 legend.title = element\_blank(),  
 #legend.position = c(0.8,0.9),  
 legend.key.size = unit(0.8,"line"),  
 plot.title=element\_text( hjust=0.00, face='bold', size=10)  
   
 )



### The Stacked Area graph

The proportional stacked Area graph is conceptually like the proportional stacked bar graph`. With the stacked area graph the object is to show both trend and proportional difference over time time for a number of data series.

The entire graph represents the total of all the data plotted.

Stacked Area Graphs also use the areas to convey whole numbers, so they do not work for negative values. Overall, they are useful for comparing multiple variables changing over an interval.

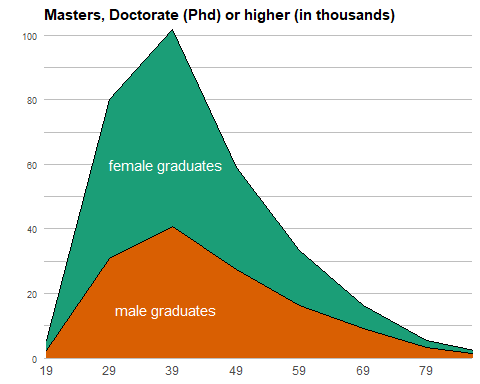
The Stacked Area graph has its standard and proportional form - analagous to the forms of the stacked bar chart

This relies upon the geom\_area geometric. It requires that the xaxis is coontinuous so we need to create a set of continuous points. I choose the mid points of the age intervals as the values for the continuous scale. We can then use the geom\_area geometric in ggplot to draw and colour an area under the line to connect the points

The stacked area graph can be misleading. If some was to read the peak number for female graduates as indicated by the y axis – you might mistakenly conclude that there are over 100,000 female graduates. However, this not true, the ridge of green is stacked on top of orange ridge so you need to substract the orange value from any value you read on the y-axis for green.

library(scales)  
  
#lower part of age range  
ire\_education$age\_continuous <- c(19,29,39,49,59,69,79,89)  
  
  
  
ggplot(ire\_education, aes(x = age\_continuous, y=number, fill=gender, colour = gender)) +  
   
geom\_area(stat="identity", colour="black", size = .2) +  
   
scale\_x\_continuous(name = "age (years)", expand = c(-0.05, 0), breaks=c(19,29,39,49,59,69,79,89), labels=c(19,29,39,49,59,69,79,89), limits = c(15,90)) +  
   
scale\_y\_continuous(name = "", breaks = seq(0,100000, by=20000), labels = seq(0,100, by=20), expand=c(0,0)) +   
   
 #scale\_fill\_brewer(palette = "Set1", name = "Uploads") +  
   
 scale\_fill\_manual(values = c( "#1b9e77", "#d95f02")) +  
 scale\_colour\_manual(values = c( "#1b9e77", "#d95f02")) +  
   
 ggtitle("Masters, Doctorate (Phd) or higher (in thousands)") +  
  
 annotate("text", x=38, y = 60000, label = "female graduates", size = 4, family = "Helvetica", colour = "white") +  
   
 annotate("text", x=38, y = 15000, label = "male graduates", size = 4, family = "Helvetica", colour = "white") +  
   
   
 theme\_classic() +  
   
 theme(  
 axis.line.x = element\_blank(),  
 axis.ticks.x = element\_blank(),  
 axis.ticks.y = element\_blank(),  
 axis.line.y= element\_blank(),  
 axis.title.x=element\_blank(),   
 axis.text.x = element\_text(size=9),  
 axis.text.y = element\_text(size=7),  
 axis.title.y = element\_text(size=9),  
 legend.text = element\_text(size=8),  
 legend.title = element\_blank(),  
 legend.position = "none",  
 legend.key.size = unit(0.8,"line"),  
 panel.grid.major.y = element\_line(size = 0.1, linetype = 'solid', colour = "grey"),  
 panel.grid.minor.y = element\_line(size = 0.1, linetype = 'solid', colour = "grey"),  
 plot.title=element\_text( hjust=0.00, face='bold', size=11)  
   
 )

## Warning in grid.Call.graphics(C\_text, as.graphicsAnnot(x$label), x$x, x$y, :  
## font family not found in Windows font database



### The Proportional Stacked Area graph

We can produce a proportion al version of this also – which visually emphasises the difference in overall proportion and the trend.

We have to be very careful using the proportional however.

While the message seems very clear it is also misleading to some extent. Because the graphic gives no indication that there are relatively few Data points in the older age groups. Without context, readers may interpret that there are as many graduates in the 25 to 34 age group as there are in the 85 and over age group. And we know that is not the case

library(scales)  
  
ire\_education$age\_continuous <- c(19,29,39,49,59,69,79,89)  
  
  
ggplot(ire\_education, aes(x = age\_continuous, y=number, fill=gender, colour = gender)) +  
   
geom\_area(position="fill", colour="black", size = .2) +  
   
scale\_x\_continuous(name = "age (years)", expand = c(0, 0), breaks=c(19,29,39,49,59,69,79,89), labels=c(19,29,39,49,59,69,79,89), limits = c(15,90)) +  
   
scale\_y\_continuous(expand = c(0, 0), labels = scales::percent, name = "") +  
  
 annotate("text", x=55, y = 0.7, label = "female graduates", size = 4, family = "Helvetica", colour = "white") +  
   
 annotate("text", x=55, y = 0.3, label = "male graduates", size = 4, family = "Helvetica", colour = "white") +  
   
   
 scale\_fill\_manual(values = c( "#1b9e77", "#d95f02")) +  
 scale\_colour\_manual(values = c( "#1b9e77", "#d95f02")) +  
   
ggtitle("Percentage Masters, Doctorate (Phd) or higher per age group") +  
   
 theme\_classic() +  
   
 theme(  
 axis.line.x = element\_blank(),  
 axis.ticks.x = element\_blank(),  
 axis.ticks.y = element\_blank(),  
 axis.line.y= element\_blank(),  
 axis.title.x=element\_blank(),   
 axis.text.x = element\_text(size=9),  
 axis.text.y = element\_text(size=7),  
 axis.title.y = element\_text(size=9),  
 legend.text = element\_text(size=8),  
 legend.title = element\_blank(),  
 legend.position = "none",  
 legend.key.size = unit(0.8,"line"),  
 plot.title=element\_text( hjust=0.00, face='bold', size=11)  
   
 )

## Warning in grid.Call.graphics(C\_text, as.graphicsAnnot(x$label), x$x, x$y, :  
## font family not found in Windows font database

