1. Correlation

install.packages("ggplot2")

library(ggplot2)

data("midwest")

midwest

### **Scatterplot**

# Scatterplot visulization

gg <- ggplot(midwest, aes(x=area, y=poptotal)) +

geom\_point(aes(col=state, size=popdensity)) +

geom\_smooth(method="loess", se=F) +

xlim(c(0, 0.1)) +

ylim(c(0, 500000)) +

labs(subtitle="Area Vs Population",

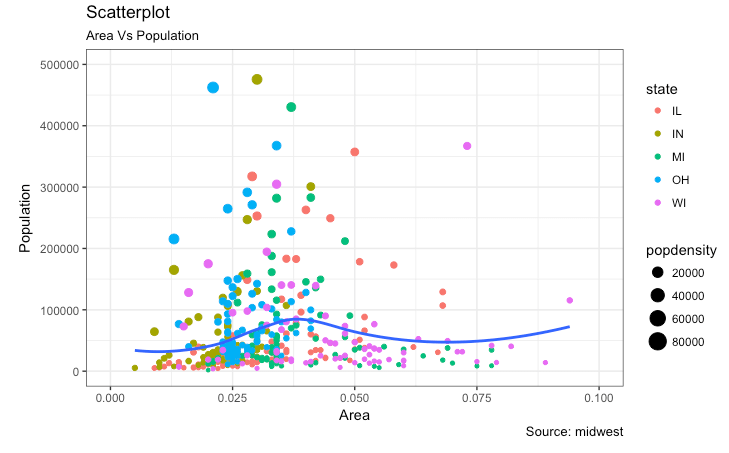
y="Population",

x="Area",

title="Scatterplot",

caption = "Source: midwest")

plot(gg)



### **Scatterplot With Encircling**

When presenting the results, sometimes I would encirlce certain special group of points or region in the chart so as to draw the attention to those peculiar cases. This can be conveniently done using the geom\_encircle() in ggalt package.

# Scatterplot With Encircling visulization

install.packages('ggalt')

library(ggalt)

# data gathering

midwest\_select <- midwest[midwest$poptotal > 350000 &

midwest$poptotal <= 500000 &

midwest$area > 0.01 &

midwest$area < 0.1, ]

# Plot

ggplot(midwest, aes(x=area, y=poptotal)) +

geom\_point(aes(col=state, size=popdensity)) + # draw points

geom\_smooth(method="loess", se=F) +

xlim(c(0, 0.1)) +

ylim(c(0, 500000)) + # draw smoothing line

geom\_encircle(aes(x=area, y=poptotal),

data=midwest\_select,

color="red",

size=2,

expand=0.08) + # encircle

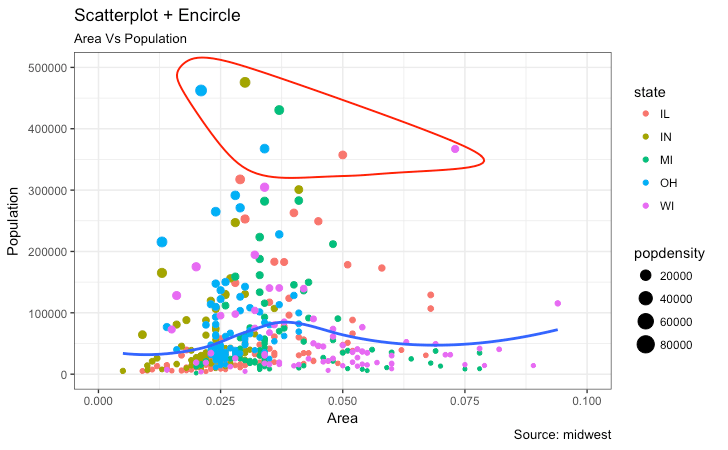
labs(subtitle="Area Vs Population",

y="Population",

x="Area",

title="Scatterplot + Encircle",

caption="Source: midwest")



### **Jitter Plot**

plot city mileage (cty) vs highway mileage (hwy).

# store data

# jitter plot visulization

data(mpg)

# store data

g <- ggplot(mpg, aes(cty, hwy))

# Scatterplot

g + geom\_point() +

geom\_smooth(method="lm", se=F) +

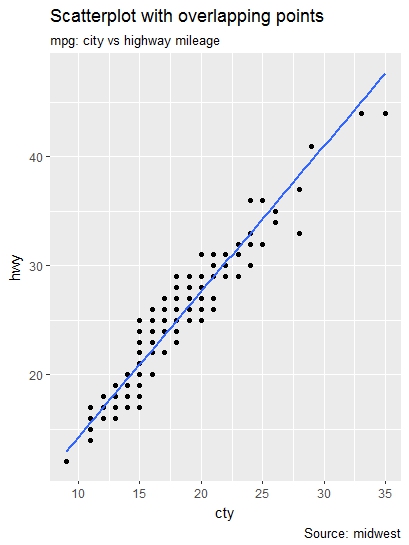
labs(subtitle="mpg: city vs highway mileage",

y="hwy",

x="cty",

title="Scatterplot with overlapping points",

caption="Source: midwest")



The original data has 234 data points but the chart seems to display fewer points. This is because there are many overlapping points appearing as a single dot.

We can make a jitter plot with jitter\_geom(). As the name suggests, the overlapping points are randomly jittered around its original position based on a threshold controlled by the width argument.

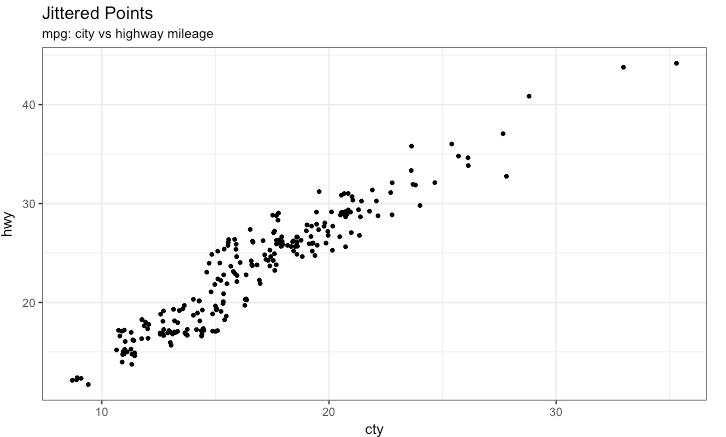
g + geom\_jitter(width = .5, size=1) +

labs(subtitle="mpg: city vs highway mileage",

y="hwy",

x="cty",

title="Jittered Points")



More the width, more the points are moved jittered from their original position.

### **Counts Chart**

The second option to overcome the problem of data points overlap is to use what is called a counts chart. Whereever there is more points overlap, the size of the circle gets bigger.

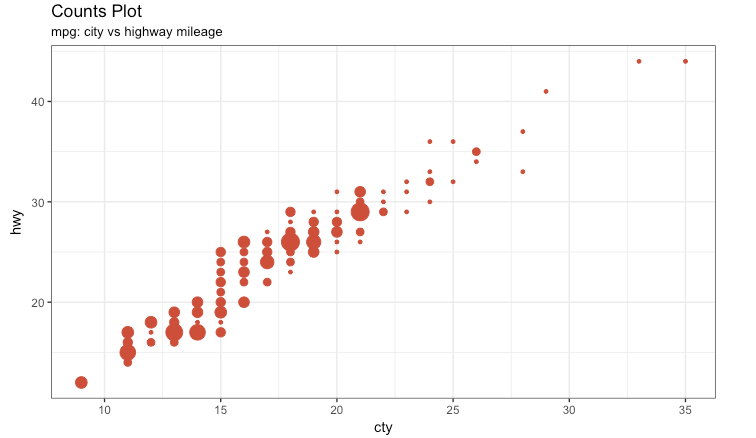
g + **geom\_count**(col="tomato3", show.legend=F) +

**labs**(subtitle="mpg: city vs highway mileage",

y="hwy",

x="cty",

title="Counts Plot")



### **Bubble plot**

While scatterplot lets you compare the relationship between 2 continuous variables, bubble chart serves well if you want to understand relationship within the underlying groups based on:

1. A Categorical variable (by changing the color) and
2. Another continuous variable (by changing the size of points).

if you have 4-Dimensional data where two of them are numeric (X and Y) and one other categorical (color) and another numeric variable (size).

mpg\_select <- mpg[mpg$manufacturer %in% c("audi", "ford", "honda", "hyundai"), ]

g <- ggplot(mpg\_select, aes(displ, cty)) +

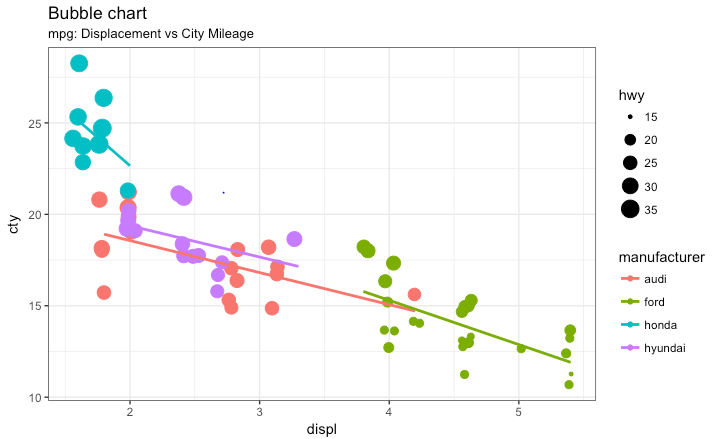
geom\_jitter(aes(col=manufacturer, size=hwy)) +

geom\_smooth(aes(col=manufacturer), method="lm", se=F) +

labs(subtitle="mpg: Displacement vs City Mileage",

title="Bubble chart")

plot(g)



### **Animated Bubble chart**

An animated bubble chart can be implemented using the gganimate package. It is same as the bubble chart, but, you have to show how the values change over a fifth dimension (typically time).

The key thing to do is to set the aes(frame) to the desired column on which you want to animate. Rest of the procedure related to plot construction is the same. Once the plot is constructed, you can animate it using gganimate() by setting a chosen interval.

# Animated Bubble chart visulization

install.packages("cowplot")

install.packages("gganimate")

library(cowplot)

library(gganimate)

library(gapminder)

View(gapminder)

g <- ggplot(gapminder, aes(gdpPercap, lifeExp, size = pop, frame = year)) +

geom\_point() +

geom\_smooth(aes(group = year),

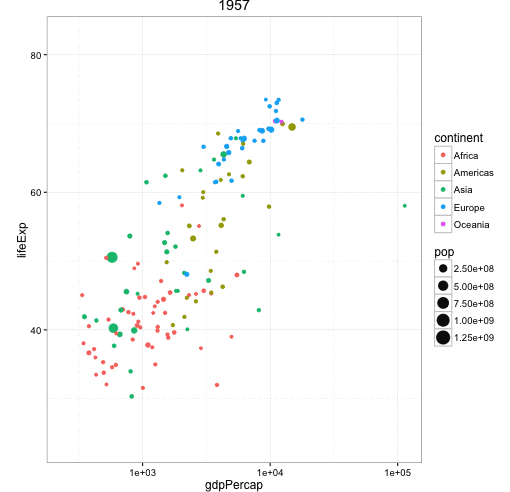
method = "lm",

show.legend = FALSE) +

facet\_wrap(~continent, scales = "free") +

scale\_x\_log10() # convert to log scale

gganimate(g, interval=0.2)



### **Marginal Histogram / Boxplot**

If you want to show the relationship as well as the distribution in the same chart, use the marginal histogram. It has a histogram of the X and Y variables at the margins of the scatterplot.

This can be implemented using the ggMarginal() function from the ‘ggExtra’ package. Apart from a histogram, you could choose to draw a marginal boxplot or density plot by setting the respective typeoption.

# Marginal Histogram / Boxplot visulization

install.packages(ggExtra)

library(ggplot2)

library(ggExtra)

library(ggMarginal)

mpg\_select <- mpg[mpg$hwy >= 35 & mpg$cty > 27, ]

g <- ggplot(mpg, aes(cty, hwy)) +

geom\_count() +

geom\_smooth(method="lm", se=F)

ggMarginal(g, type = "histogram", fill="transparent")

ggMarginal(g, type = "boxplot", fill="transparent")

# ggMarginal(g, type = "density", fill="transparent")

