Statistical Computing with R: Masters in Data Sciences 503 (S17) First Batch, SMS, TU, 2021

Shital Bhandary

Associate Professor

Statistics/Bio-statistics, Demography and Public Health Informatics

Patan Academy of Health Sciences, Lalitpur, Nepal

Faculty, Data Analysis and Decision Modeling, MBA, Pokhara University, Nepal

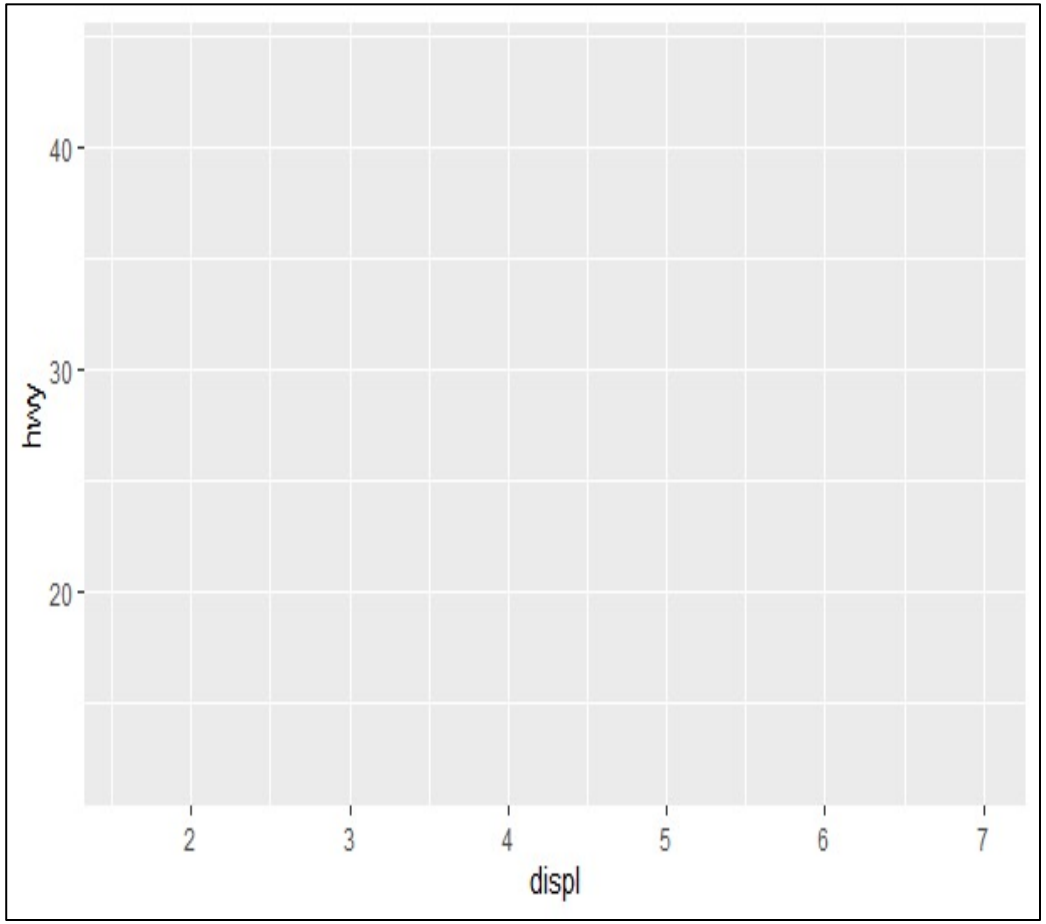
Faculty, FAIMER Fellowship in Health Professions Education, India/USA.

Review Preview

- ggplot2 package
  - Codes and use
- Summary measures
  - Codes and use

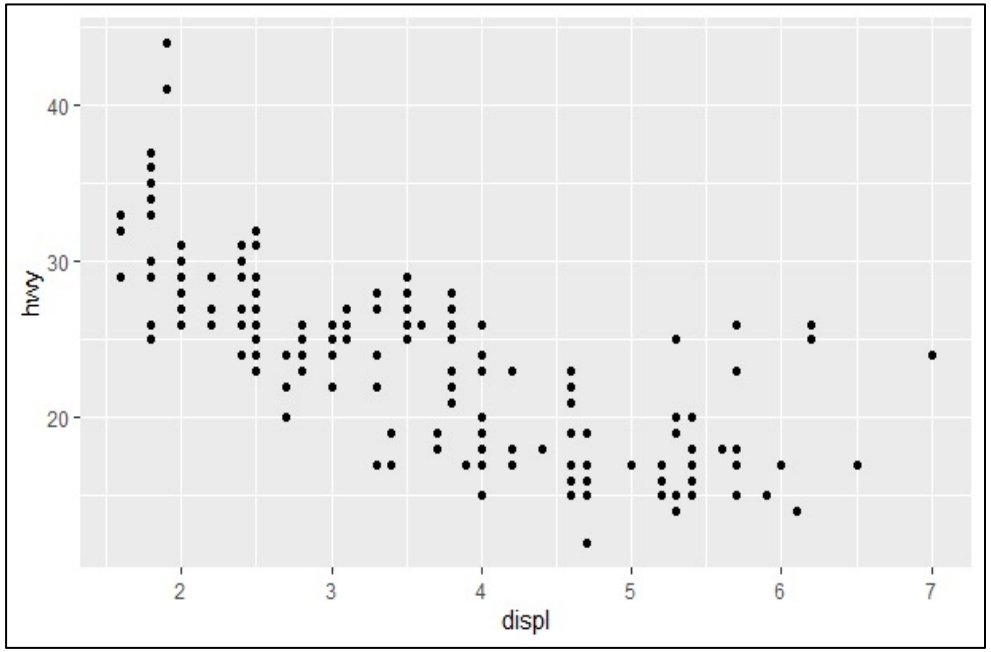
What will happen? Why?

```
#Simple point
ggplot(mpg, aes(x = displ, y = hwy))
geom_point()
```



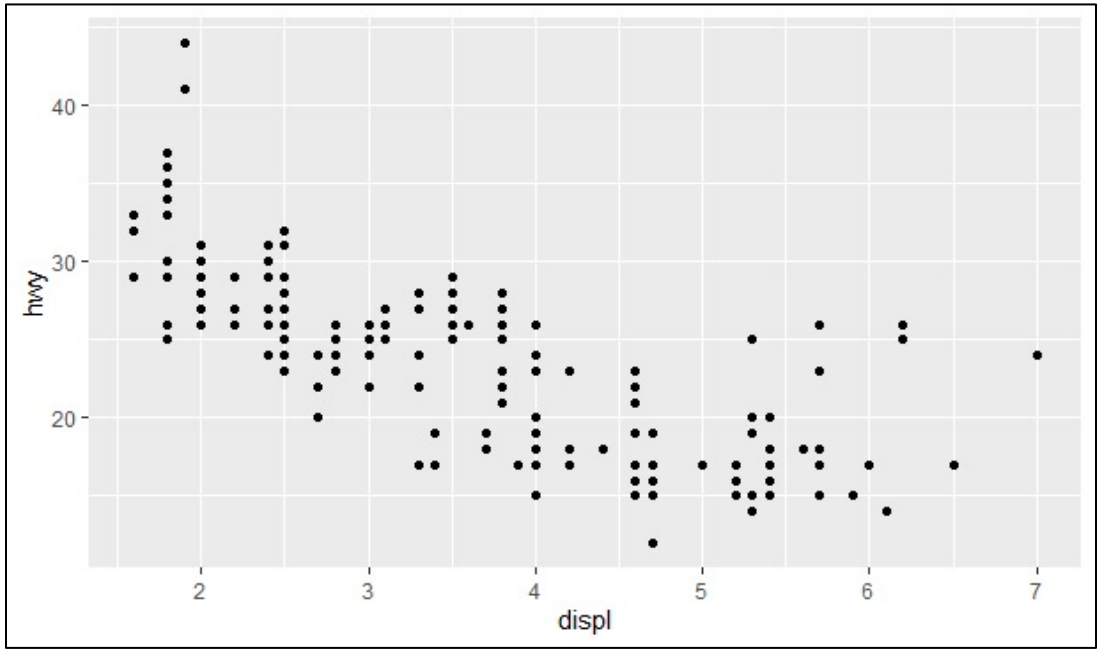
What will happen? Why?

```
#Simple point
ggplot(mpg, aes(x = displ, y = hwy)
) +
geom_point()
```



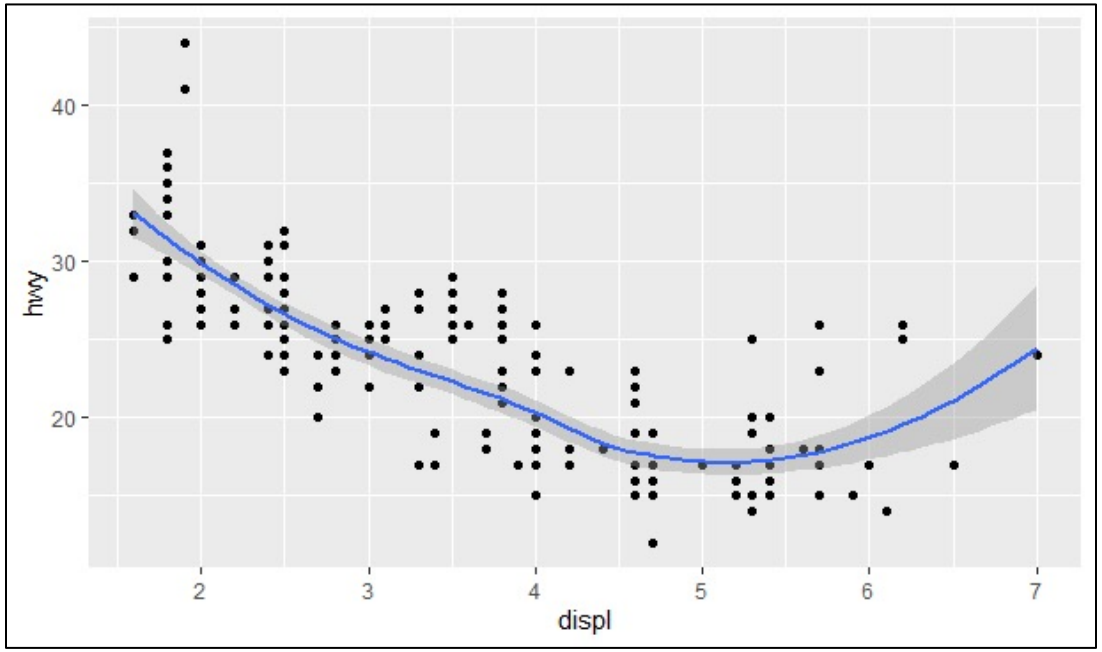
What will happen? Why?

```
#Equivalent code
ggplot(mpg, aes(displ, hwy)) +
geom_point()
```



How to add a “smooth” plot on top of this?

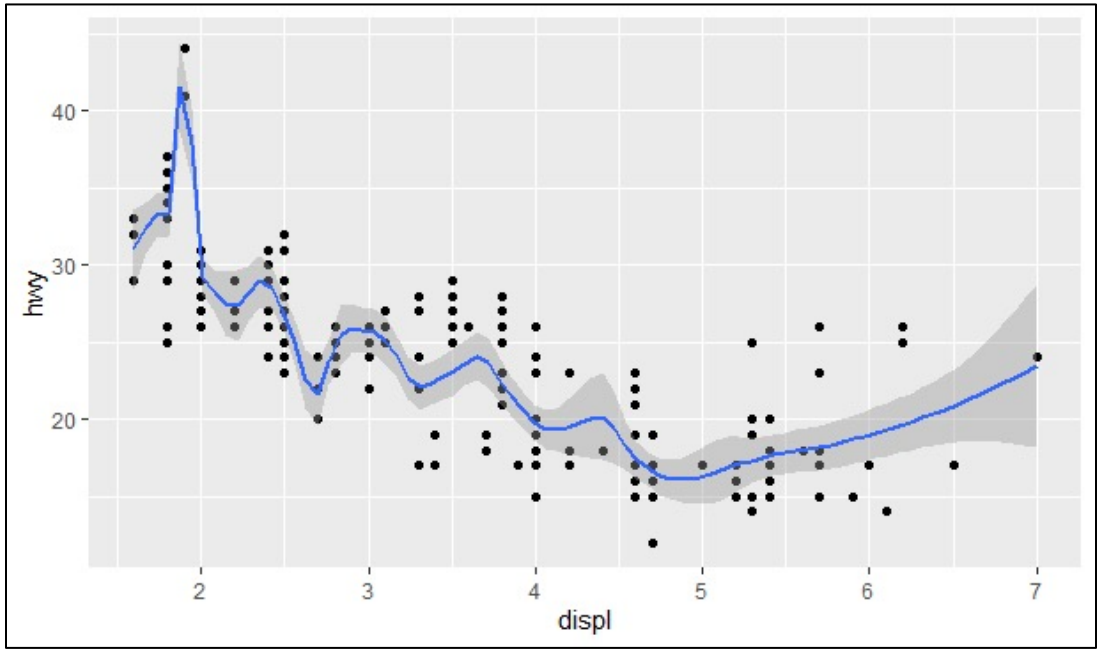
```
#Scatterplot
ggplot(mpg, aes(displ, hwy)) +
geom_point()

#Add smooth line
geom_smooth()     #??
```



Adding “wiggliness” in the smoothing plot:

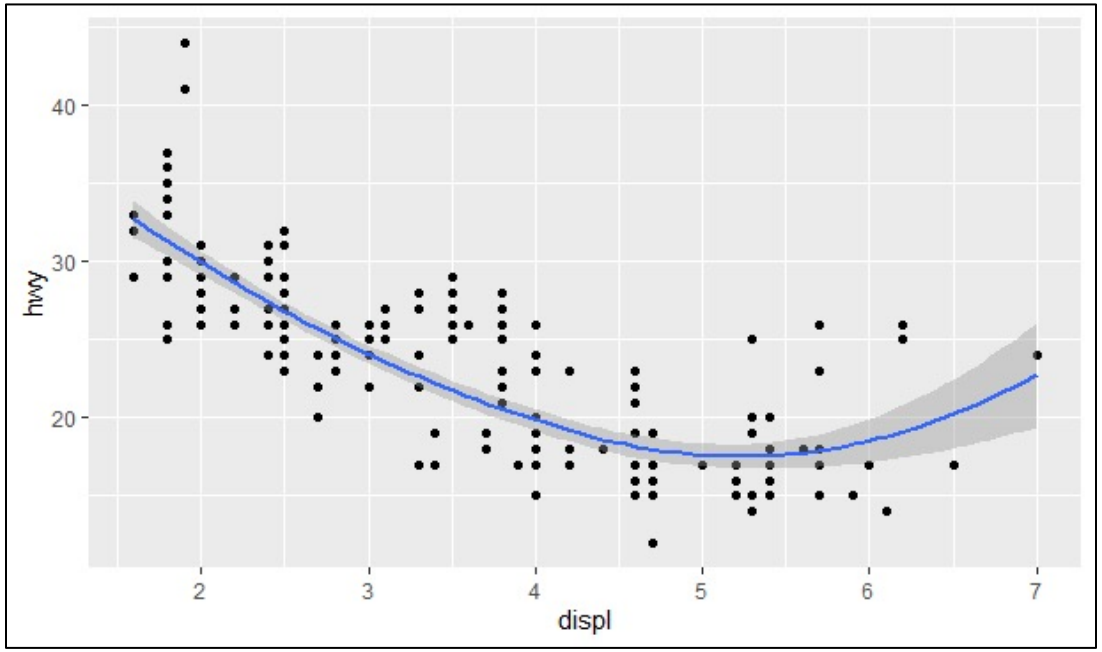
```
ggplot(mpg, aes(displ, hwy)) +
geom_point() +
geom_smooth(span = 0.2)
```

```
#Extreme wiggliness = 0.2
```



Adding “wiggliness” in the smoothing plot:

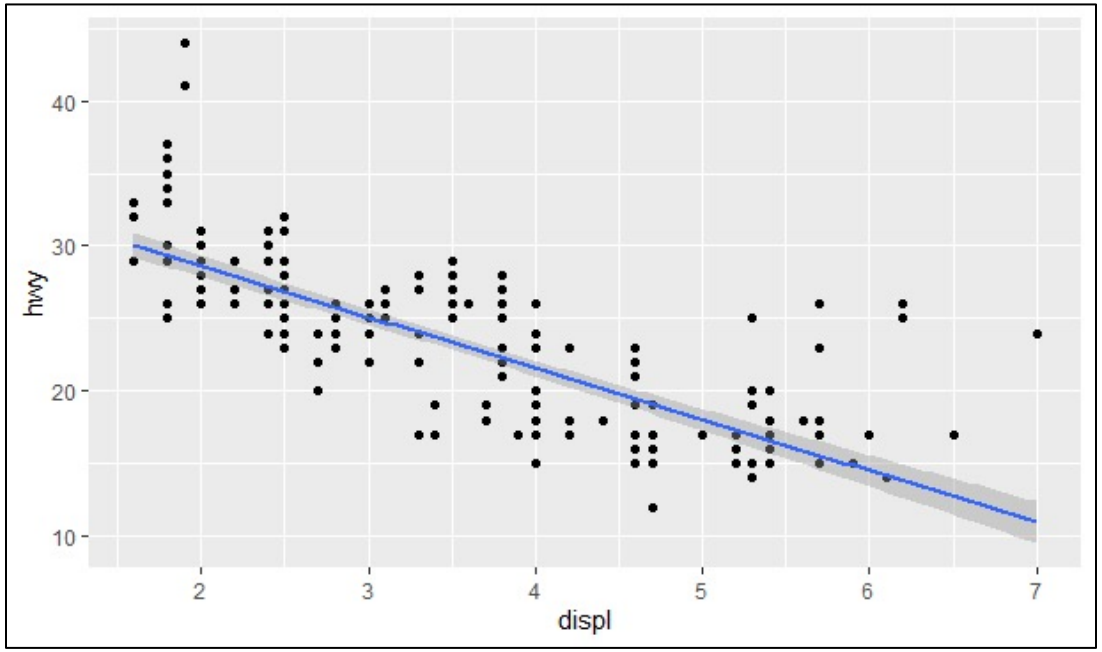
```
ggplot(mpg, aes(displ, hwy)) +
geom_point() +
geom_smooth(span = 1)
```



#No wiggliness = 1

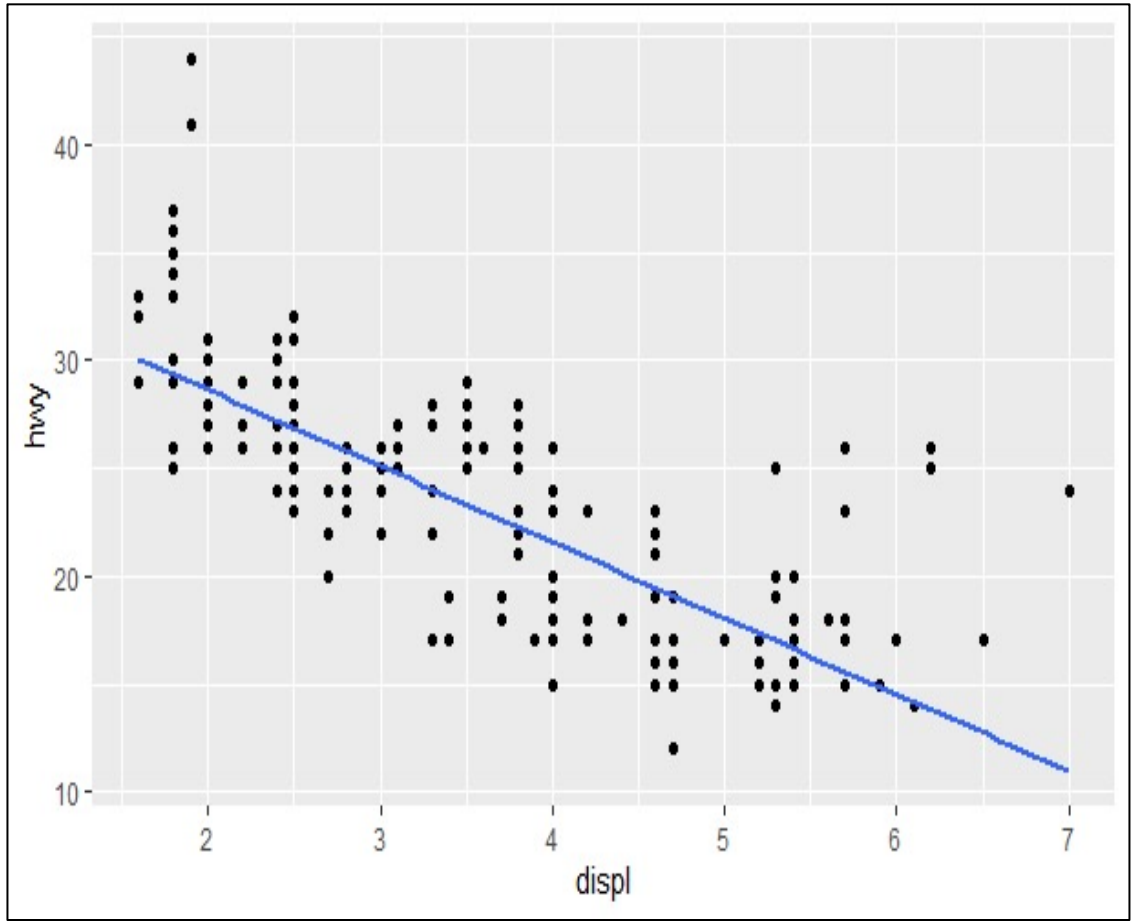
How to add a linear “smooth” plot on top of the scatterplot?

```
#Scatterplot
ggplot(mpg, aes(displ, hwy)) +
geom_point()

#Add smooth line
geom_smooth(method = lm)
```



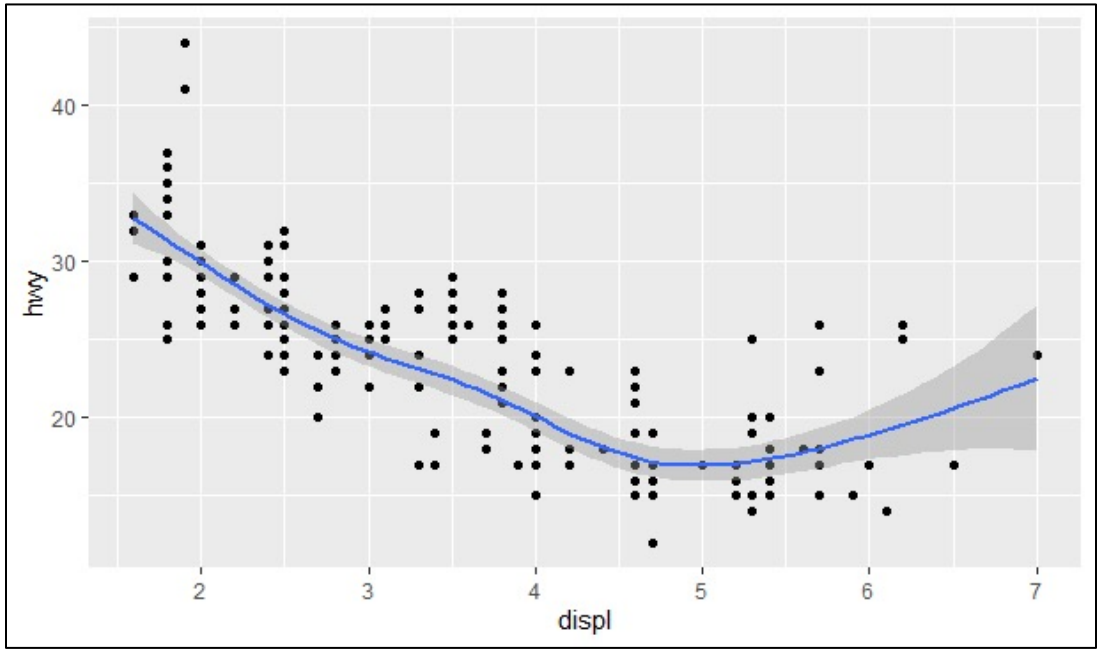
How to add a linear “smooth” plot on top of this without the “standard error” band?

```
#Scatterplot
ggplot(mpg, aes(displ, hwy)) +
geom_point()

#Add smooth line
geom_smooth(method = lm)
geom_smooth(method = lm, se = FALSE)
```



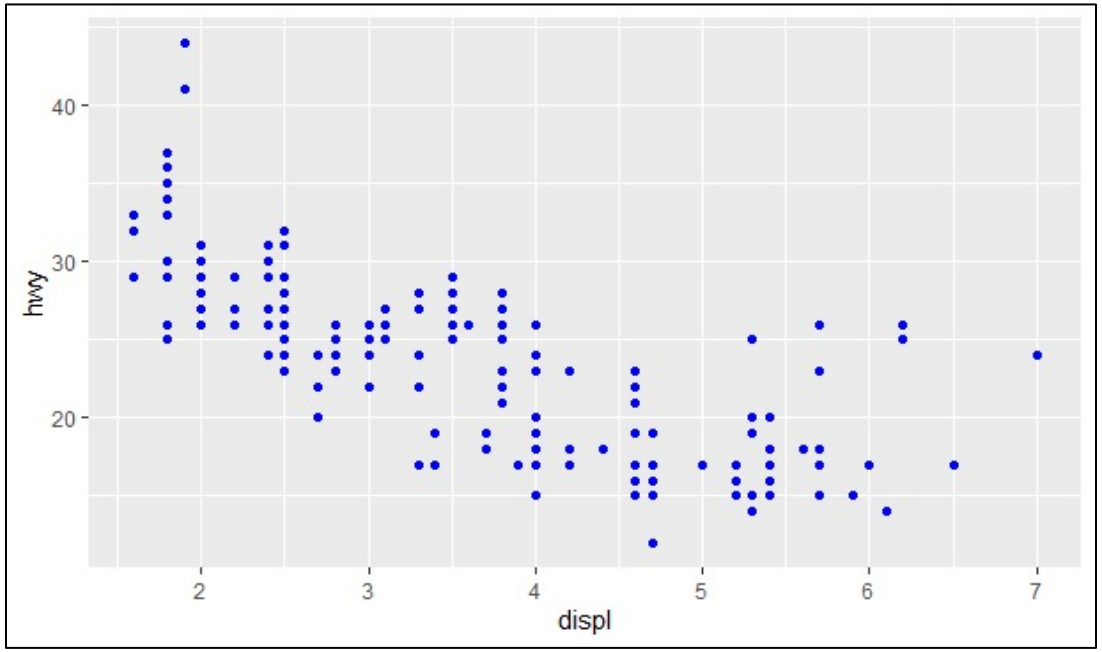
“Loess” does not work with large dataset (n > 1000)! What to do then?

```
# when n is greater than 1,000.
# we need to use “mgcv” library
# and call general additive model
# smoothing
library(mgcv)
ggplot(mpg, aes(displ, hwy)) +
geom_point() +
geom_smooth(method = "gam", formula = y ~ s(x))
```



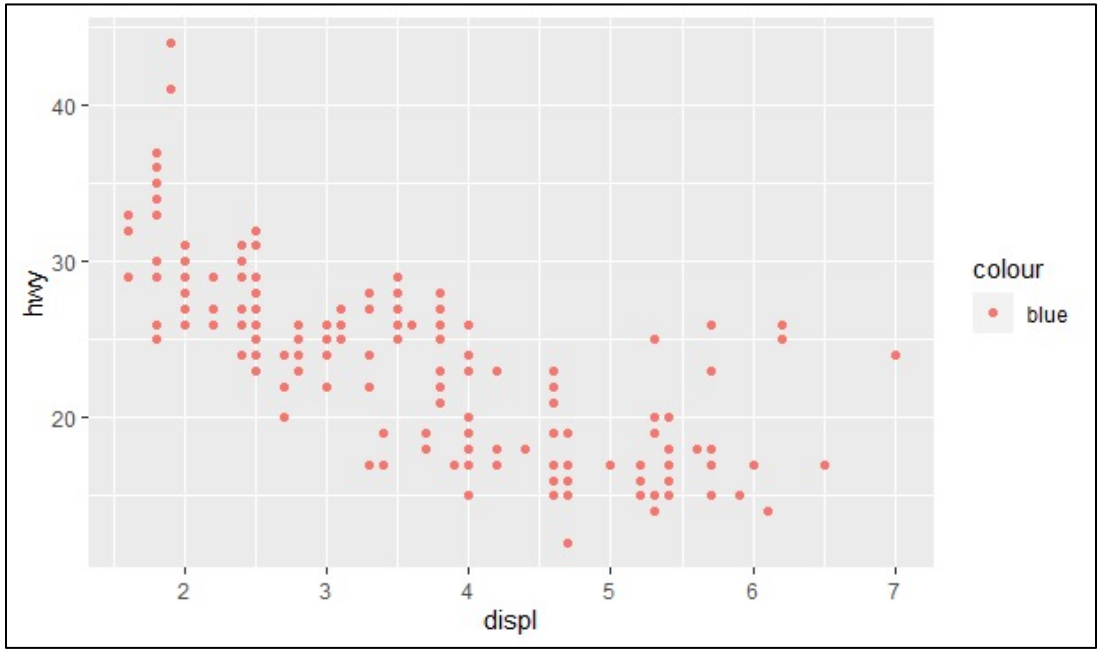
Fixed color (of points) without legend: good?

```
#Fixed color without legend
ggplot(mpg, aes(displ, hwy)) +
geom_point(colour = "blue")
```



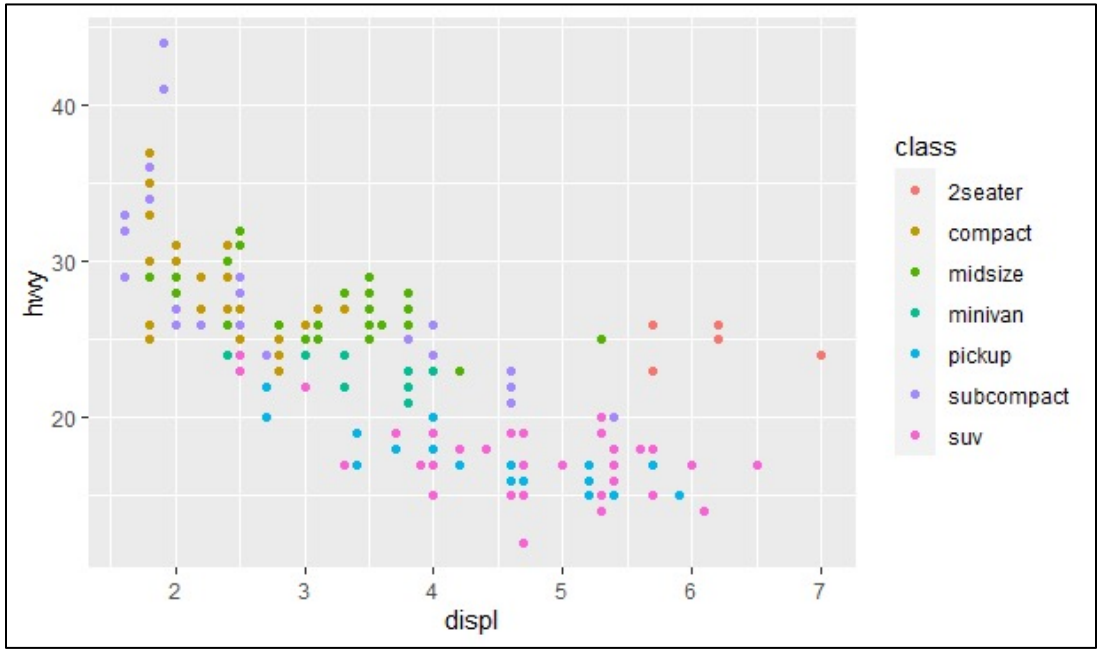
Fixed color (of points) with legend: good?

```
#Fixed color with legend
ggplot(mpg, aes(displ, hwy)) +
geom_point(aes(colour = "blue"))
```



Changing color by variable attributes:

```
#Color, size, shape and other aesthetic attributes

ggplot(mpg, aes(displ, hwy, colour = class)) +

geom_point()
```



Getting multiple scatterplot of attributes:

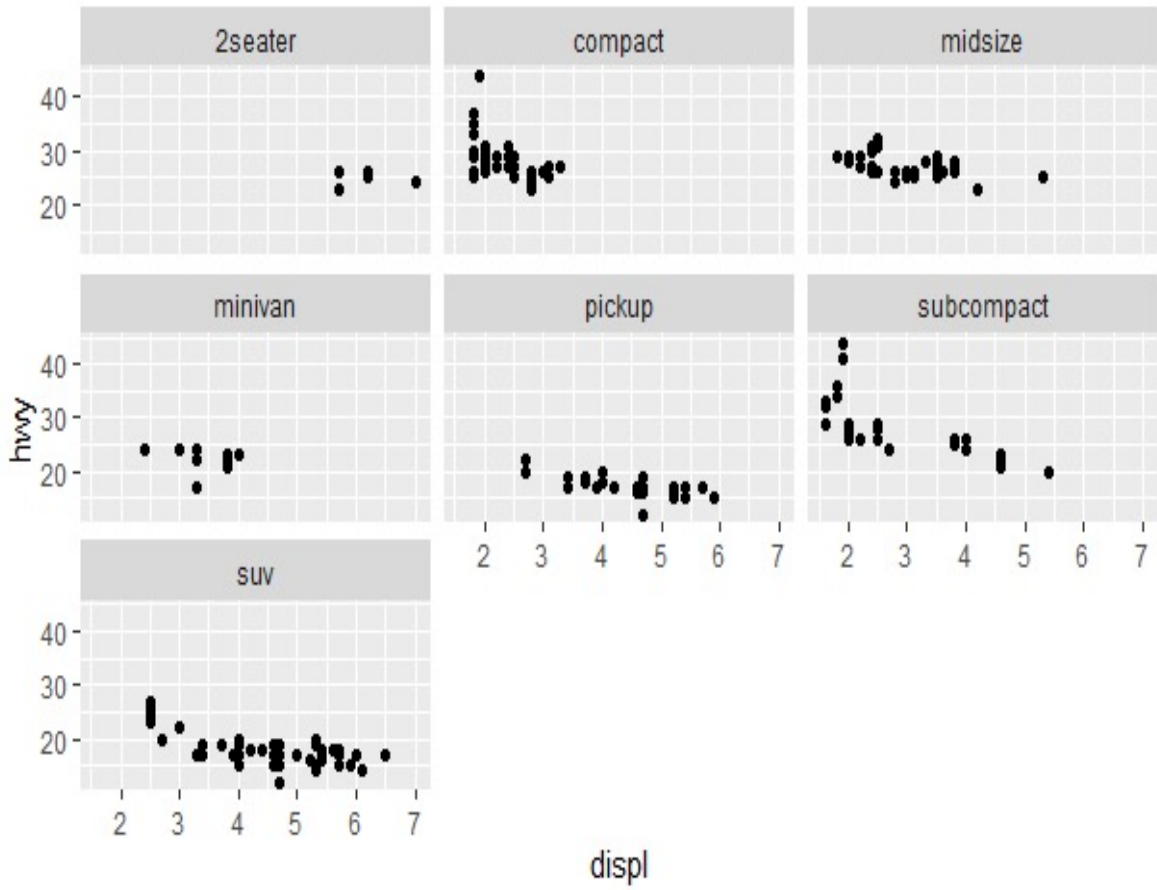
- #Faceting

ggplot(mpg, aes(displ, hwy)) +

geom_point() +

facet_wrap(~class)

#Number of columns for the graph is determined automatically, if not specified!

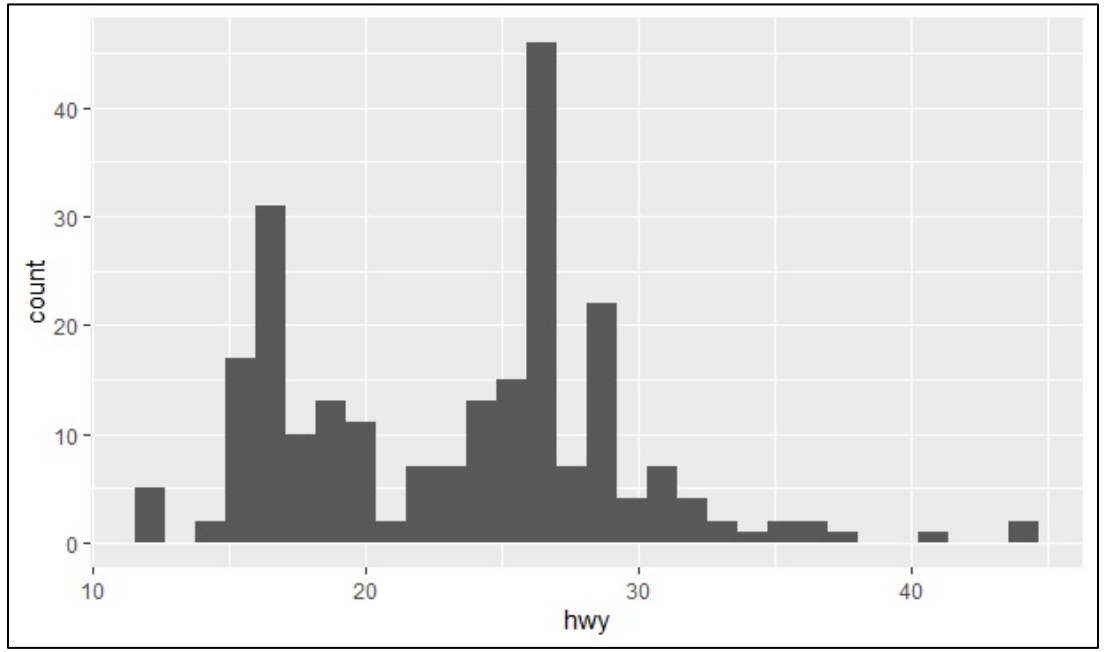


Histogram:

```
ggplot(mpg, aes(hwy)) +
  geom_histogram()
```

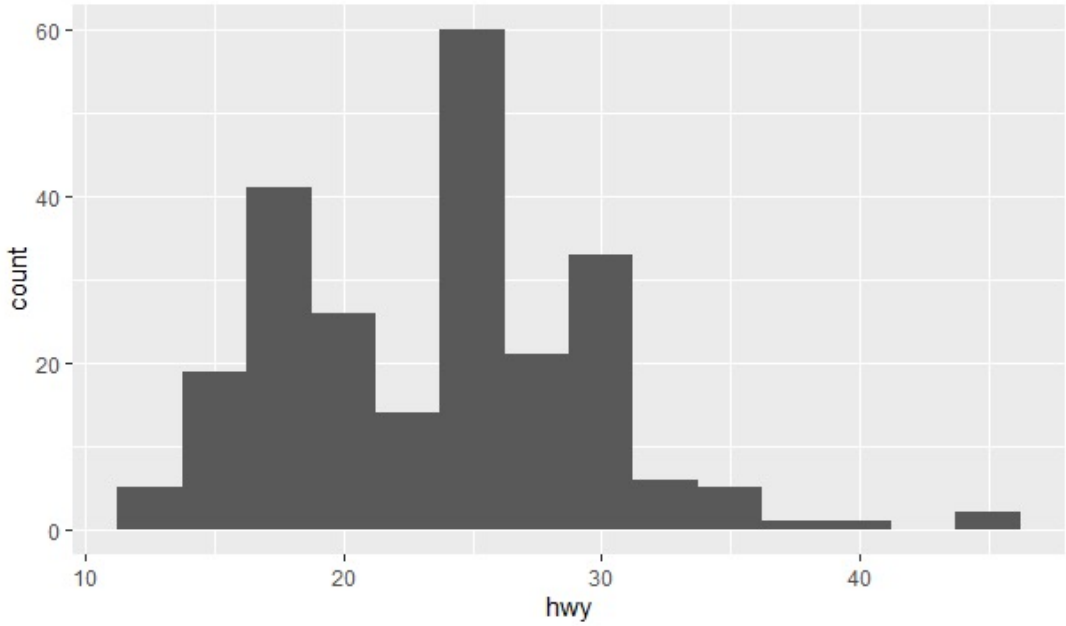
#How was it created?

hwy variable was binned "automatically"!



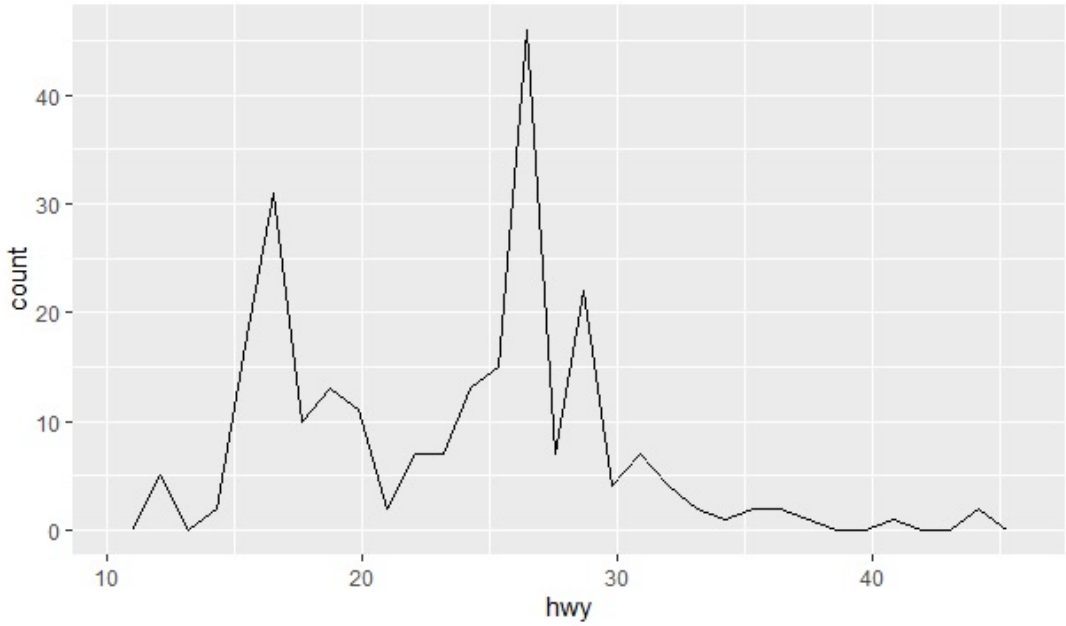
Changing bin size of the histogram:

```
ggplot(mpg, aes(hwy)) +
geom_histogram(binwidth = 2.5)
```



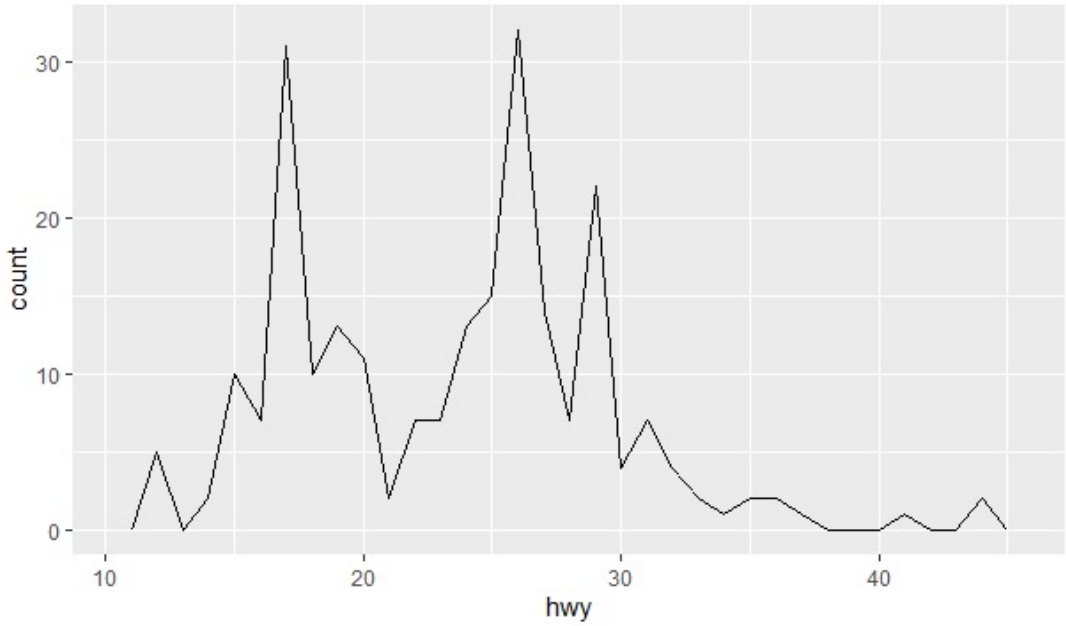
Frequency polygon:

```
ggplot(mpg, aes(hwy)) +
geom_freqpoly()

# How as it created?
```



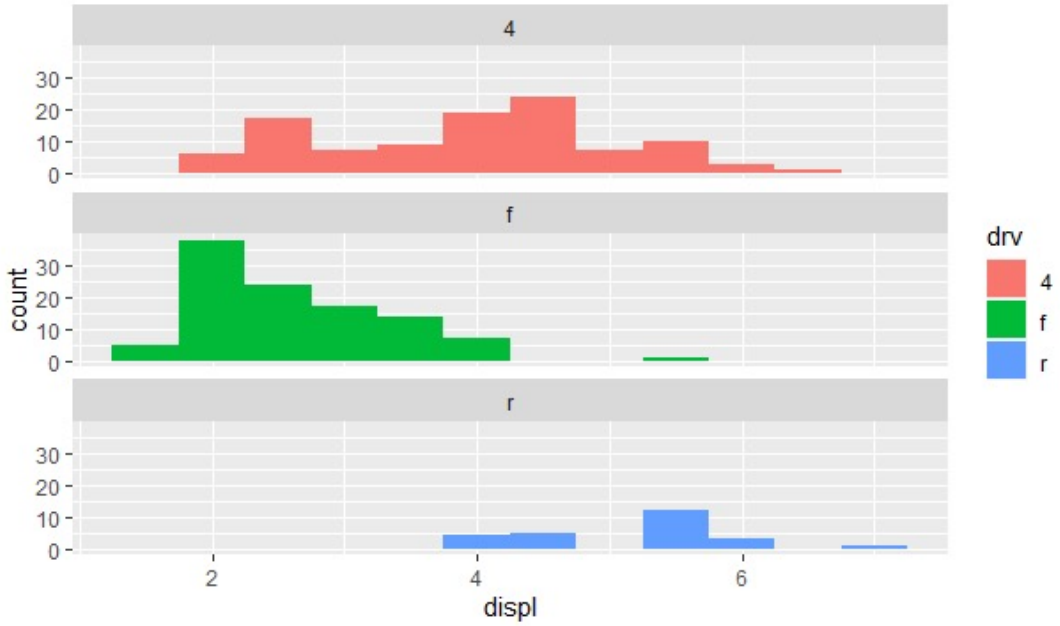
Changing bin size of frequency polygon:

```
ggplot(mpg, aes(hwy)) +
geom_freqpoly(binwidth = 1)
```



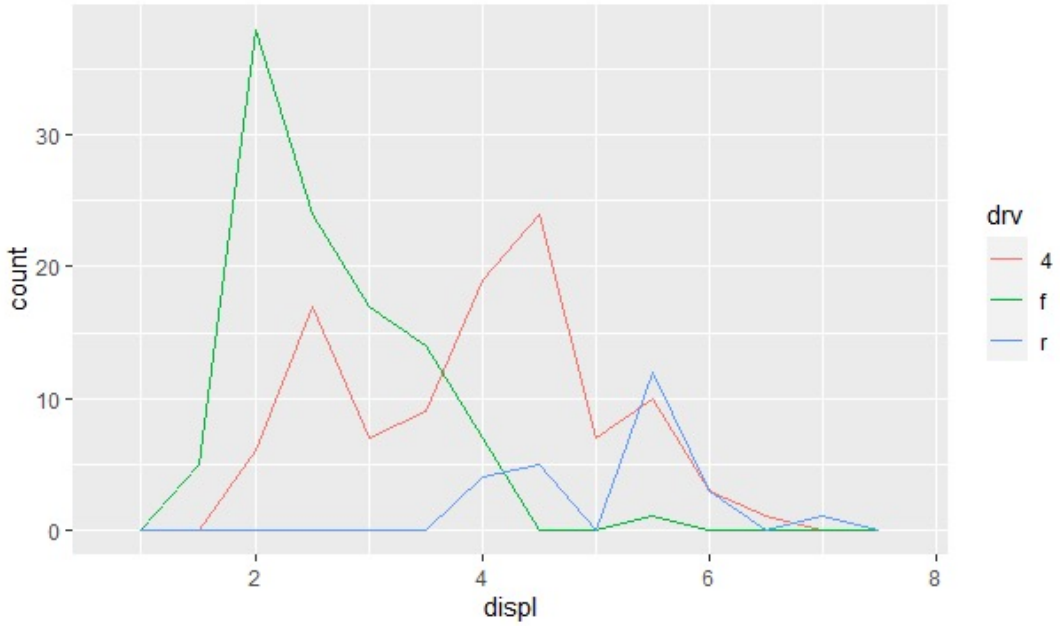
Histogram with faceting:

```
ggplot(mpg, aes(displ, fill = drv)) +
geom_histogram(binwidth = 0.5) +
facet_wrap(~drv, ncol = 1)
```



Frequency polygon with color faceting:

```
ggplot(mpg, aes(displ, colour =
drv)) +
geom_freqpoly(binwidth = 0.5)
```

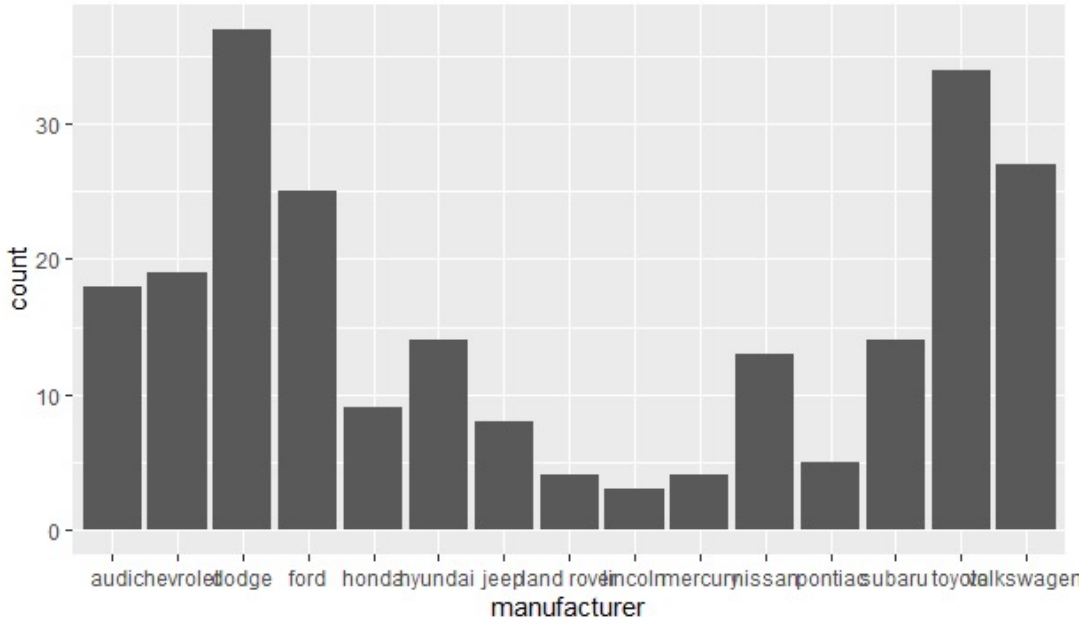


Bar plot

#Bar charts with raw data

ggplot(mpg, aes(manufacturer)) +

geom_bar()



Bar chart from frequency distribution:

```
#Bar chart with frequency distribution
drugs <- data.frame(
drug = c("a", "b", "c"),
effect = c(4.2, 9.7, 6.1)
)
```

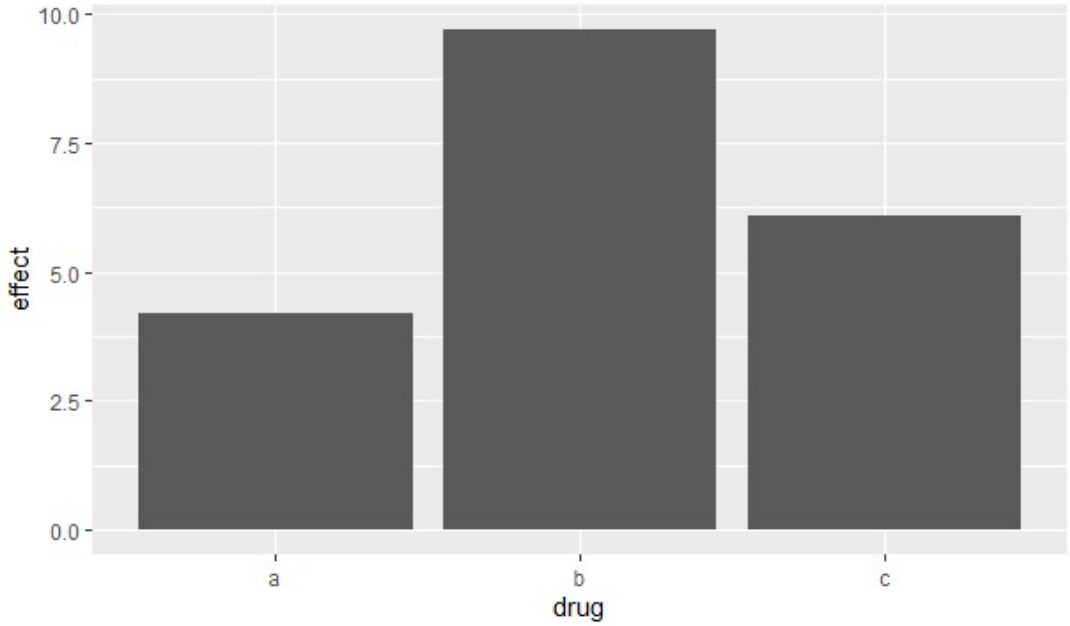
Filter

	drug	effect
1	a	4.2
2	b	9.7
3	c	6.1

Bar plot and point plot

- #Bar chart

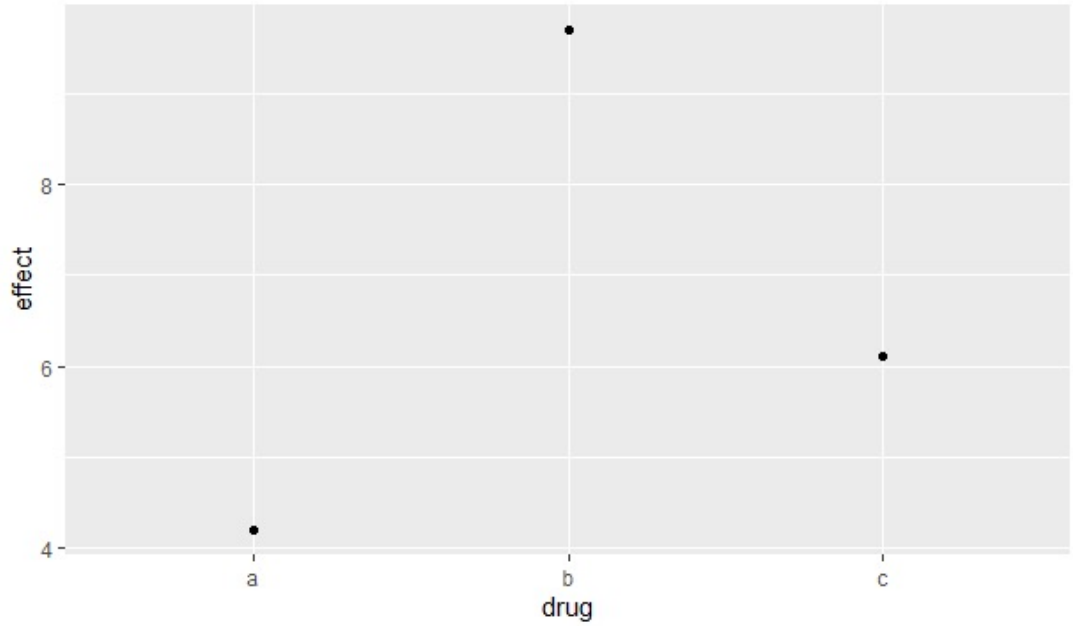
ggplot(drugs, aes(drug, effect)) +
geom_bar(stat = "identity")



Bar plot and point plot

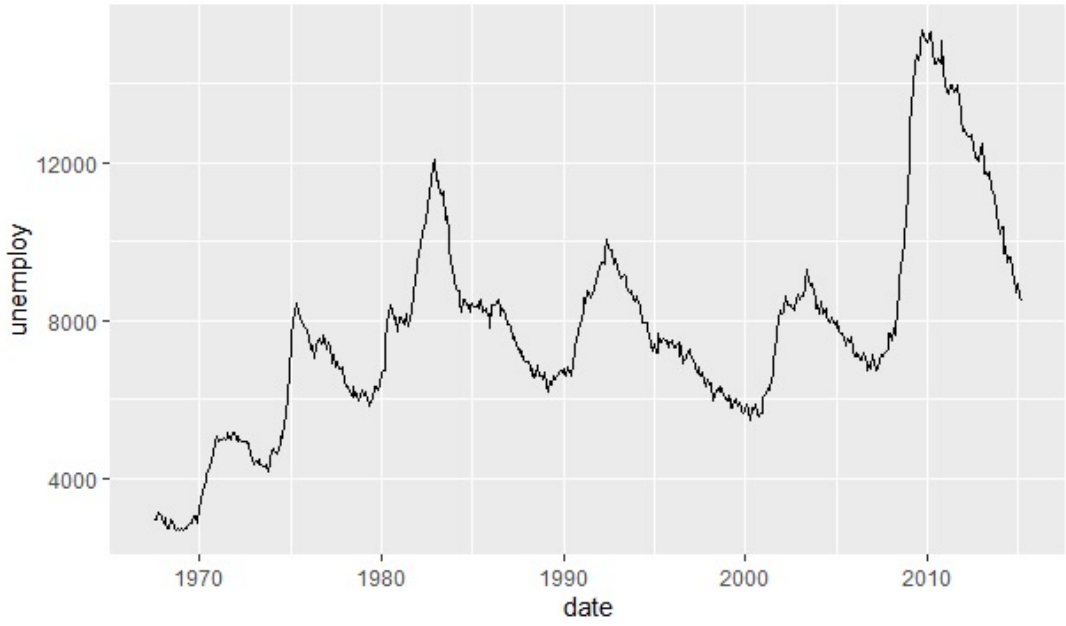
#Point chart

ggplot(drugs, aes(drug, effect)) +
geom_point()



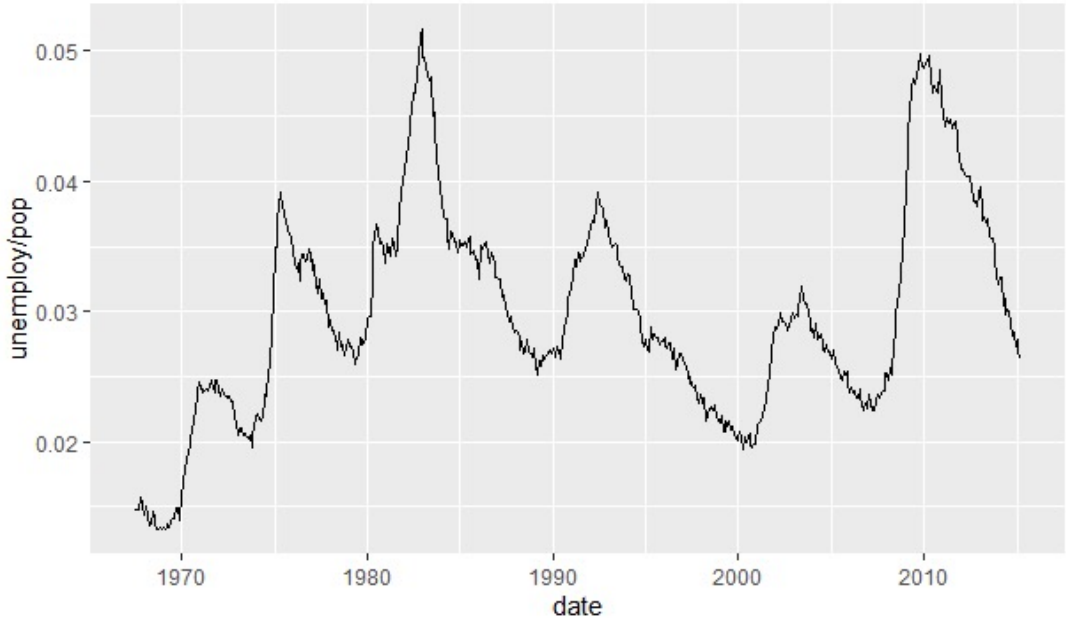
Line plot

```
#Line
ggplot(economics, aes(date, unemploy)) +
geom_line()
```



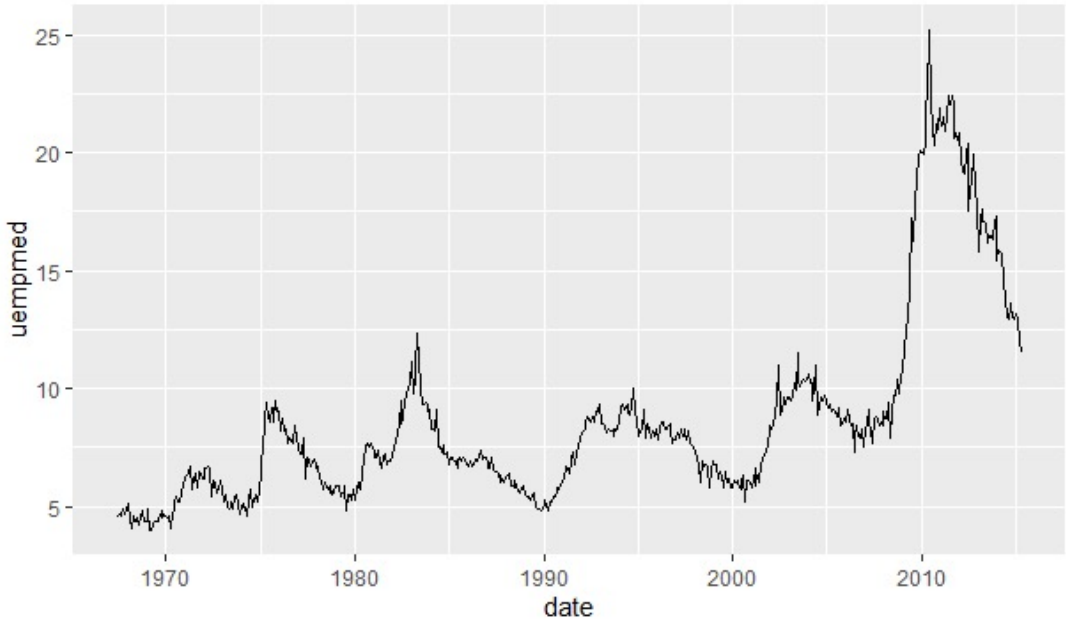
Time series plot:

```
#Time series
ggplot(economics, aes(date, unemploy / pop)) +
geom_line()
```



Time series plot:

```
#Time series
ggplot(economics, aes(date, uempmed)) +
 geom_line()
```

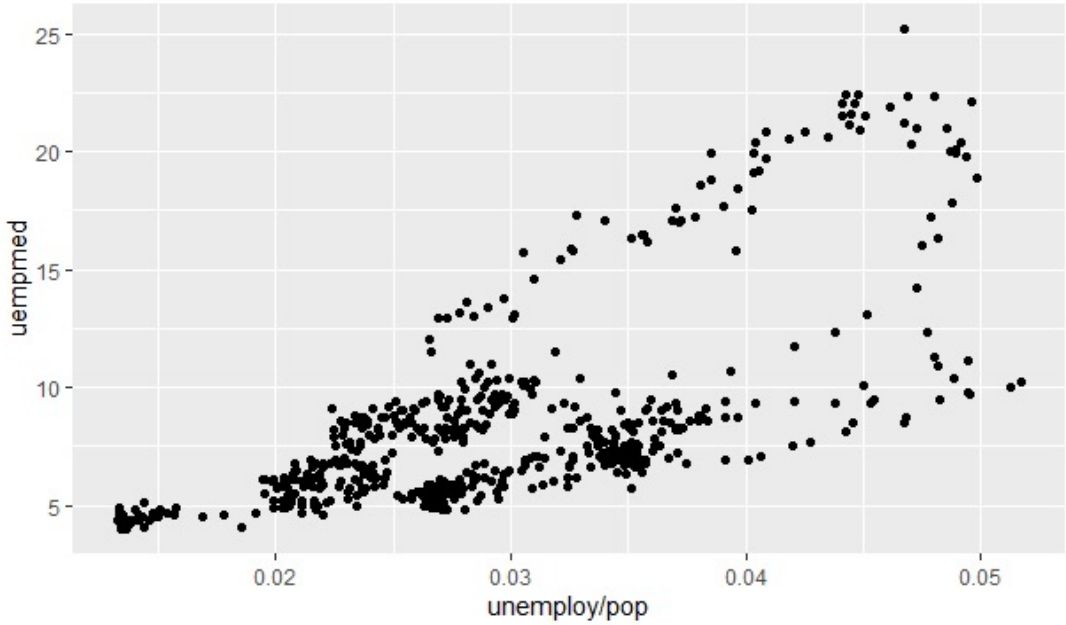


Combining Time series charts:

#Get both the charts in a single graph with scatterplot

ggplot(economics, aes(unemploy / pop, uempmed)) +

geom_point()



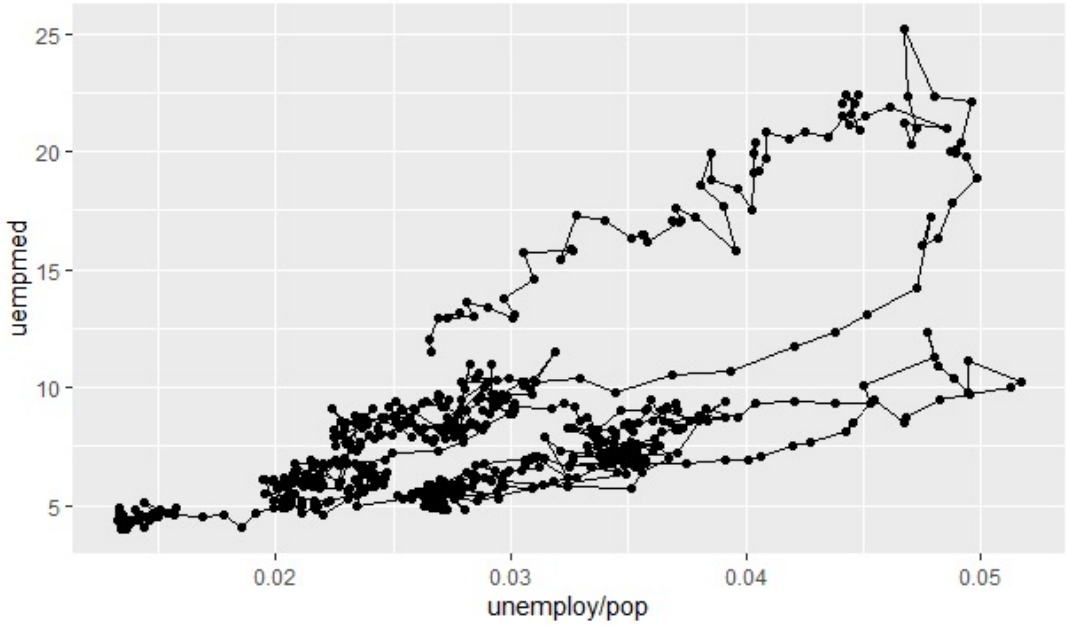
Adding paths:

#Get both the charts in a single graph with scatterplot and path

ggplot(economics, aes(unemploy / pop, uempmed)) +

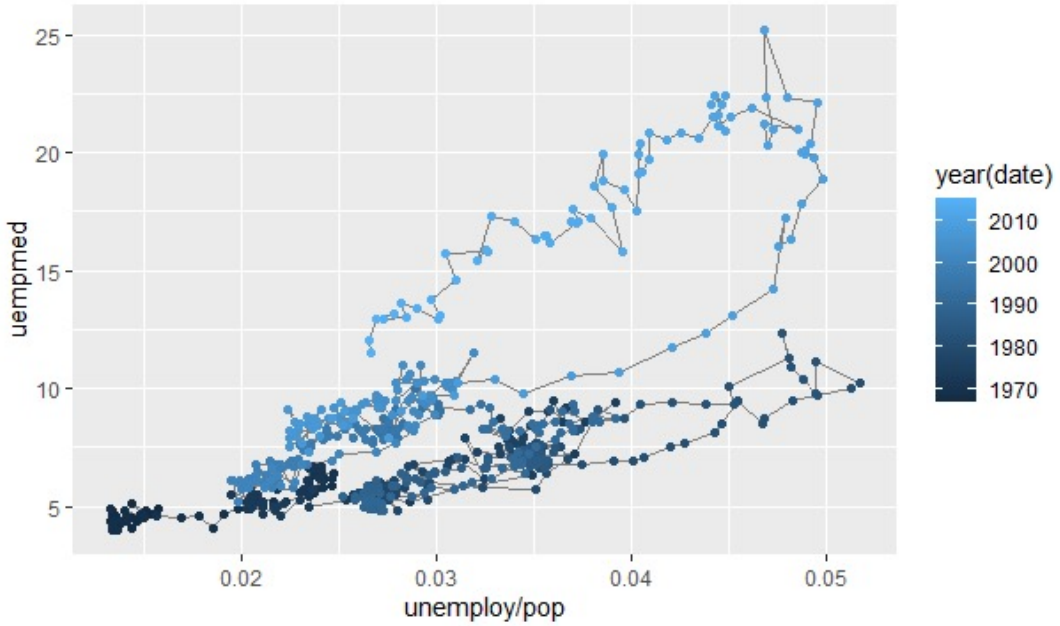
geom_path() +

geom_point()



Time series plot with time legend:

```
#Time series plot with time legend
year <- function(x)
as.POSIXlt(x)$year + 1900
ggplot(economics, aes(unemploy /
pop, uempmed)) +
geom_path(colour = "grey50") +
geom_point(aes(colour =
year(date)))
```

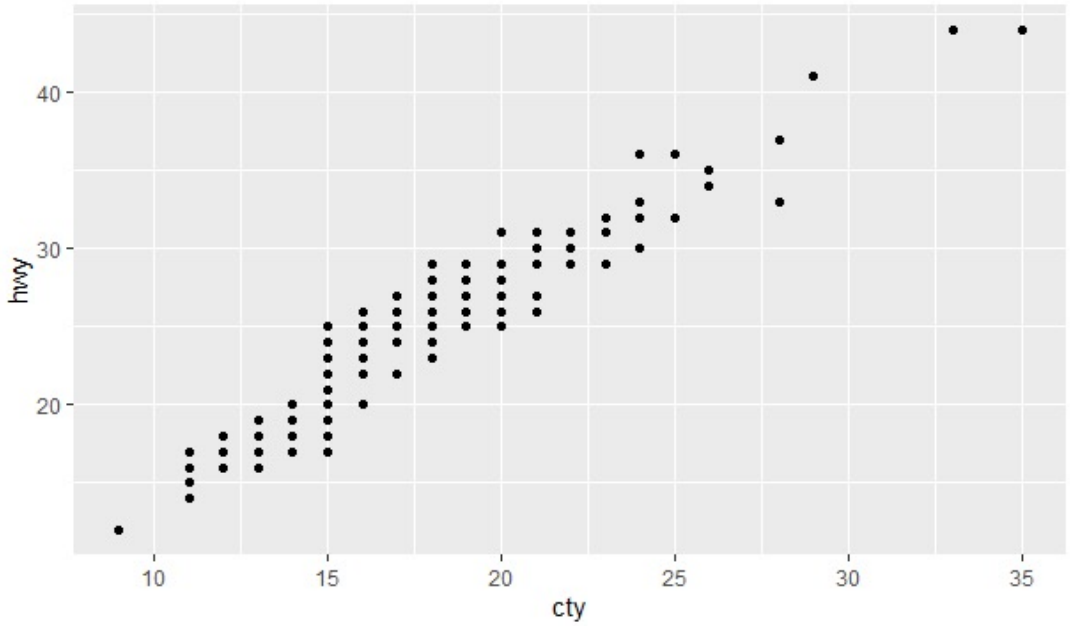


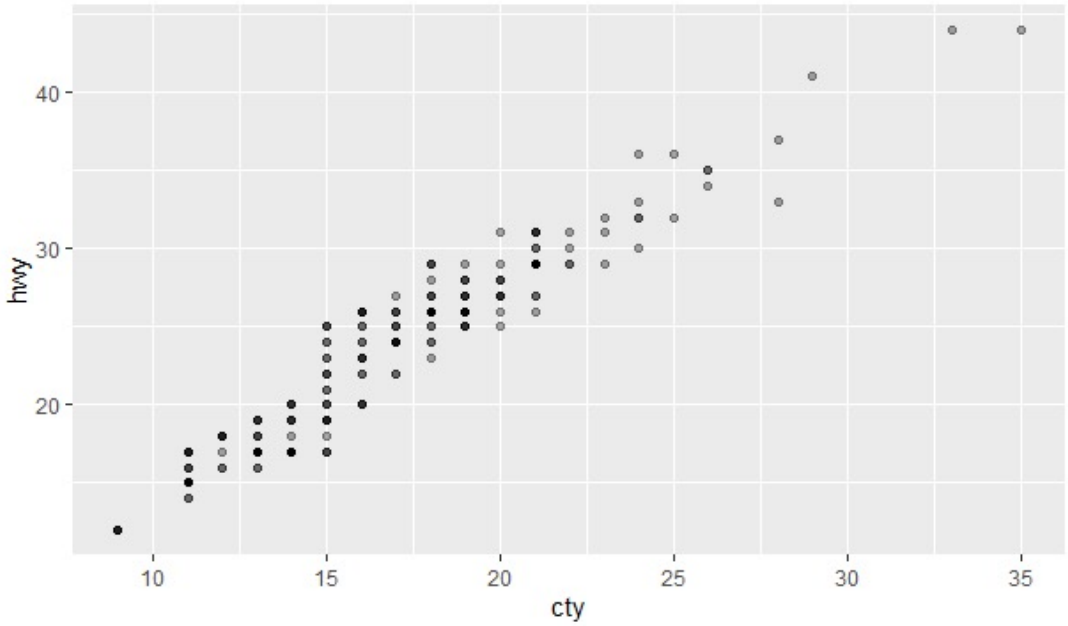
Modifying axes:

- #Scatterplot
- ggplot(mpg, aes(cty, hwy)) +
- geom_point()

- #Modifying the axes
- ggplot(mpg, aes(cty, hwy)) +
- geom_point(alpha = 1 / 3)

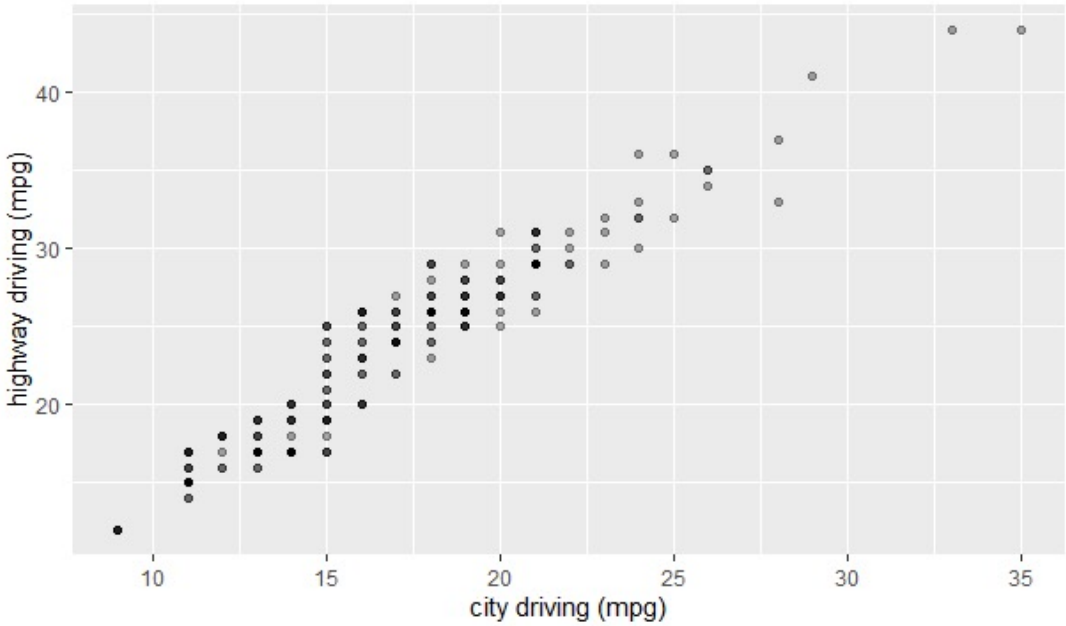
Outputs:





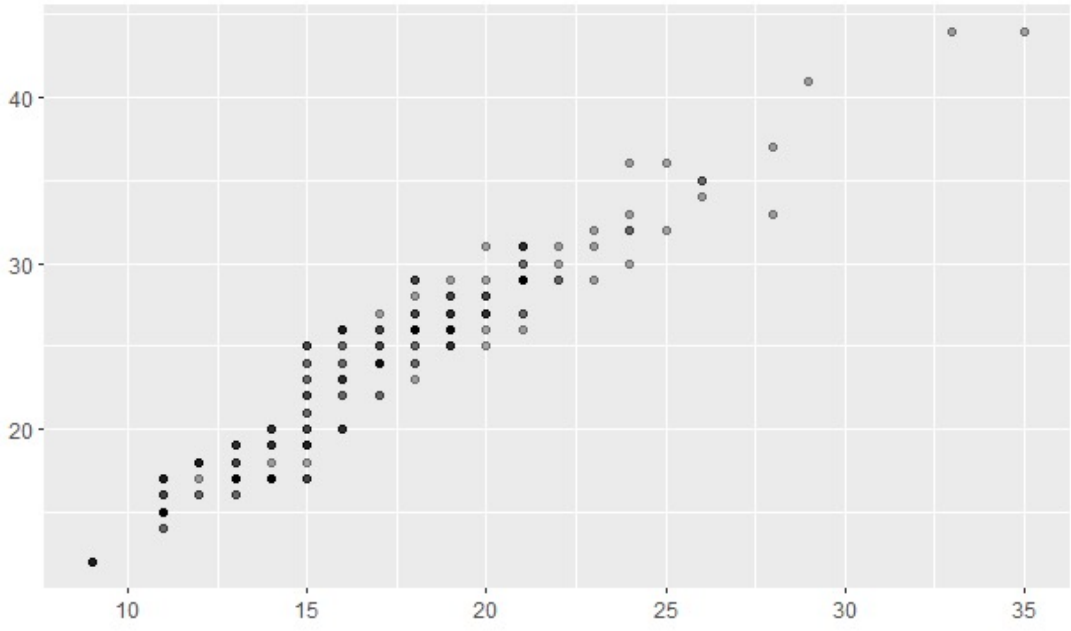
Modifying the axes:

```
#Modifying the axes
ggplot(mpg, aes(cty, hwy)) +
geom_point(alpha = 1 / 3) +
xlab("city driving (mpg)") +
ylab("highway driving (mpg)")
```



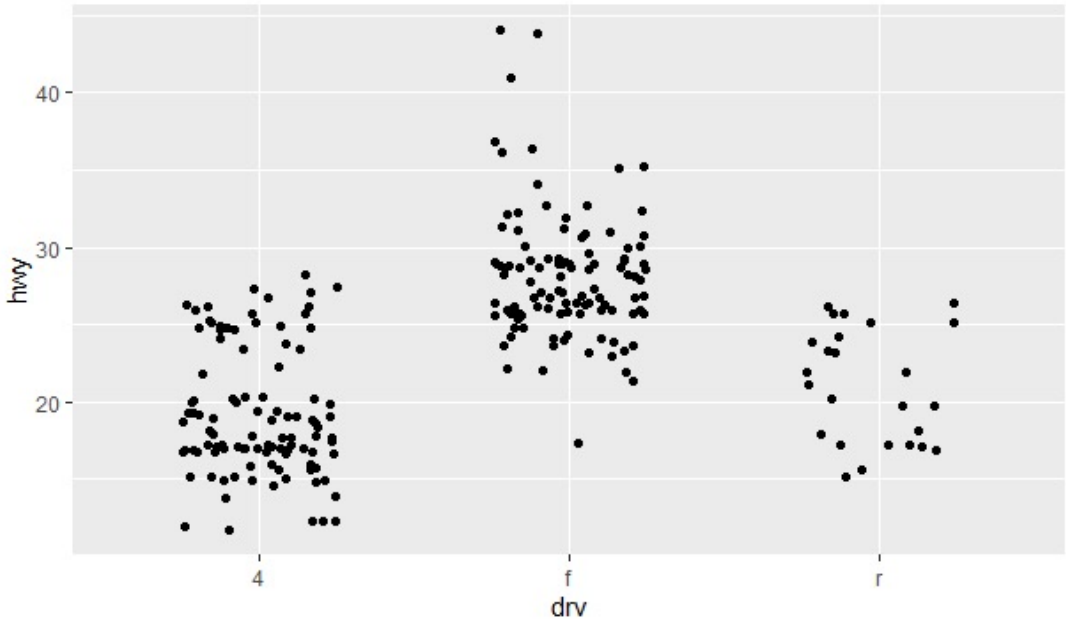
Modifying the axes:

- # Remove the axis labels with NULL
- ggplot(mpg, aes(cty, hwy)) +
- geom_point(alpha = 1 / 3) +
- xlab(NULL) +
- ylab(NULL)



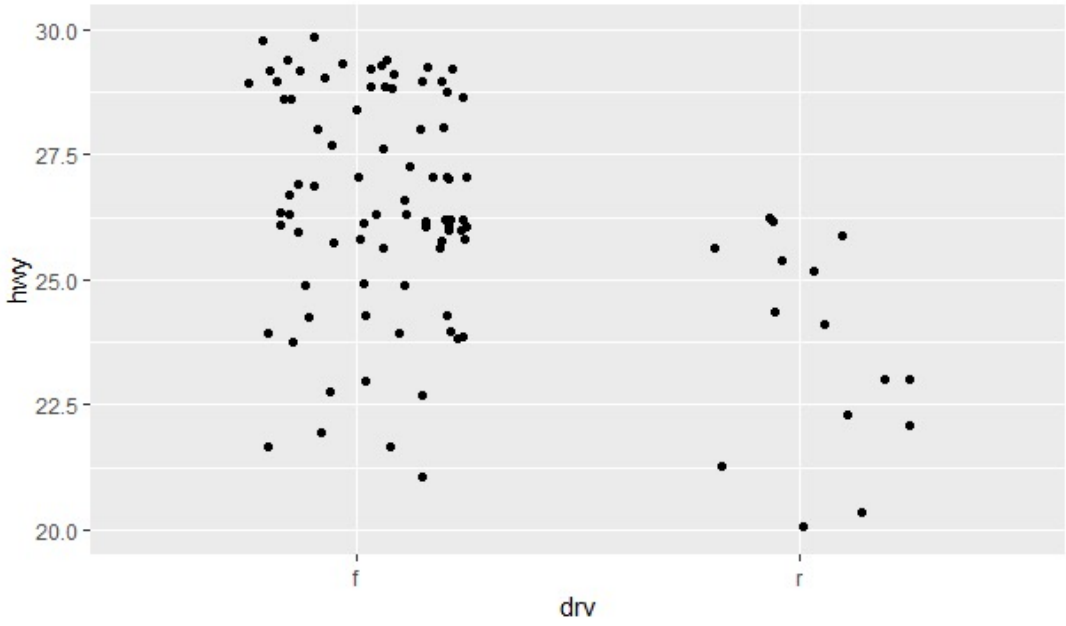
Modifying the axes:

- #Modify x and y axis limits
- ggplot(mpg, aes(drv, hwy)) +
- geom_jitter(width = 0.25)



Modifying the axes:

```
#Modify x and y axis limits
ggplot(mpg, aes(drv, hwy)) +
geom_jitter(width = 0.25) +
xlim("f", "r") +
ylim(20, 30)
```



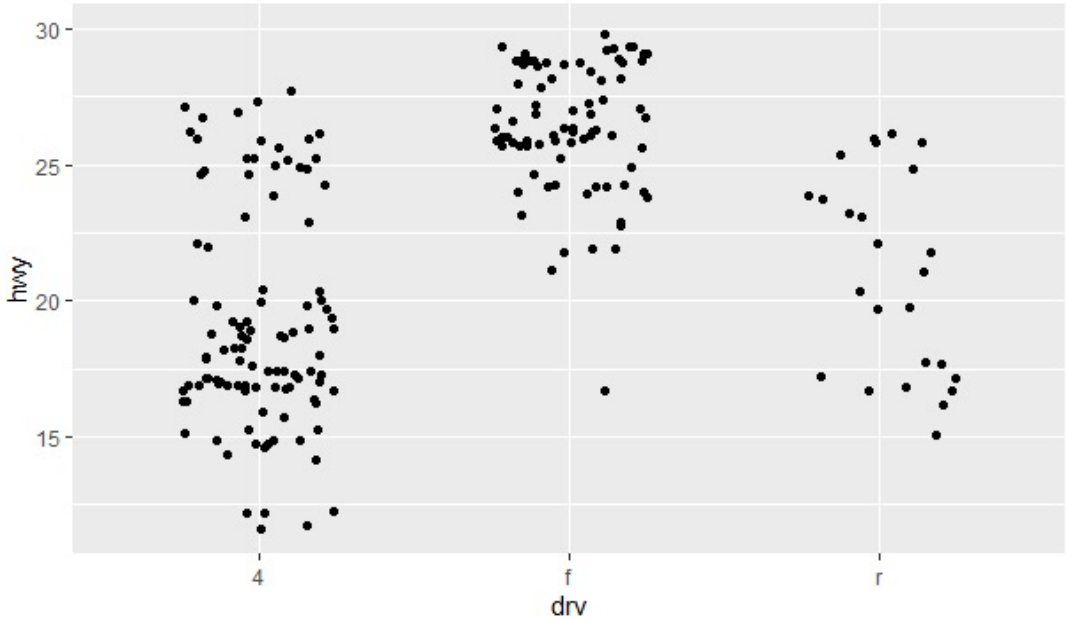
Modifying the axes:

For continuous scales, use NA to set only one limit

ggplot(mpg, aes(drv, hwy)) +

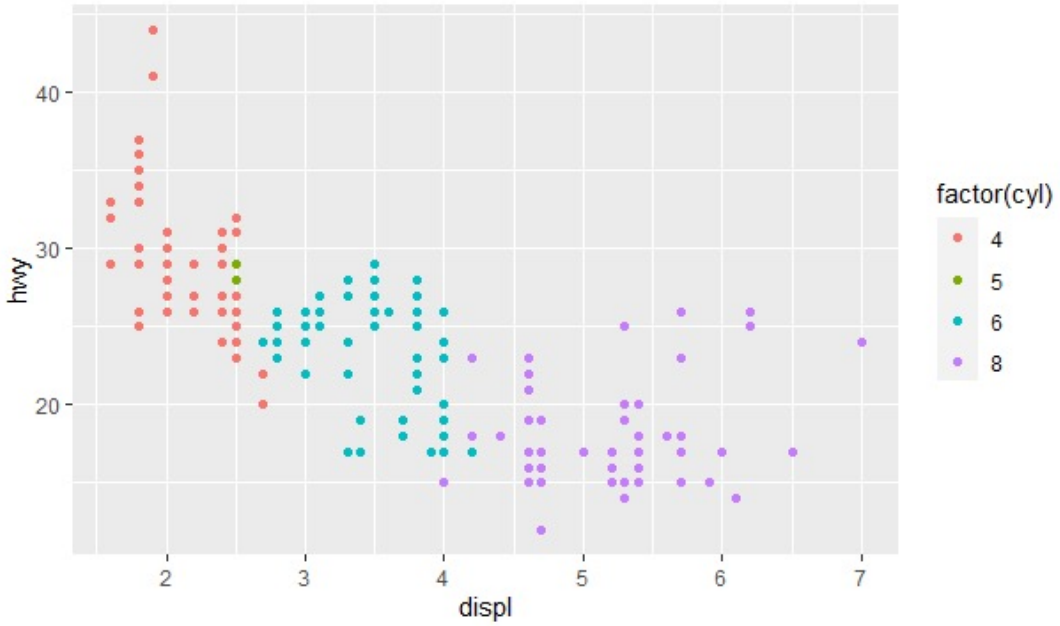
geom_jitter(width = 0.25, na.rm = TRUE) +

ylim(NA, 30)



Defining graph as an object:

```
#Graph as an object
p <- ggplot(mpg, aes(displ, hwy, colour = factor(cyl))) +
geom_point()

#Print the graph
print(p)
```



Saving the graph:

```
# Save png to disk
ggsave("plot.png", p, width = 5, height = 5)
```

- # You can change the type, width and height of the graph while saving it in the disk
- This comes handy when you need to create and submit graphs with specified “dpi” values!

Summary statistics: Review

- Exploratory data analysis
  - ~~Stem-leaf plot~~
  - Q-Q plot
  - Five number summary
  - Box-plot
    - With outliers
    - With extreme values

Summary statistics: Review

- Descriptive statistics

- Histogram
- Mean & standard deviation
- Median & Interquartile range
- Mode and range
- ~~Test of normality (Skewness and Kurtosis)~~

Data Mining vs Text Mining?

- Data Mining
  - Correlation

- Text Mining
  - Corpus
  - Pre-processing
  - Term Document Matrix
  - Most frequent terms
  - Co-occurrence of terms
  - Cluster Analysis
    - Hierarchical
    - K-means
  - Topic Modelling

Question/Queries?

- If not, review these highlighted topics (Unit 1, 2 and 3) of the course syllabus on “Data sciences” for the exam:

  - Program
  - **Import**
  - **Tidy**
  - **Visualize**
  - Model
  - Communicate

Thank you!

@shitalbhandary