Continuous + Categorical

Prof. Joyce Robbins

Graphical Data Analysis with R

3. Examining Continuous Variables
4. Displaying Categorical Data
5. Looking for Structure: Dependency Relationships and Associations
6. Investingating Multivariate Continuous Data
7. Studying Multivariate Categorical Data

Graphical Data Analysis with R

3. Examining Continuous Variables

4. Displaying Categorical Data

5. Looking for Structure: Dependency Relationships and Associations

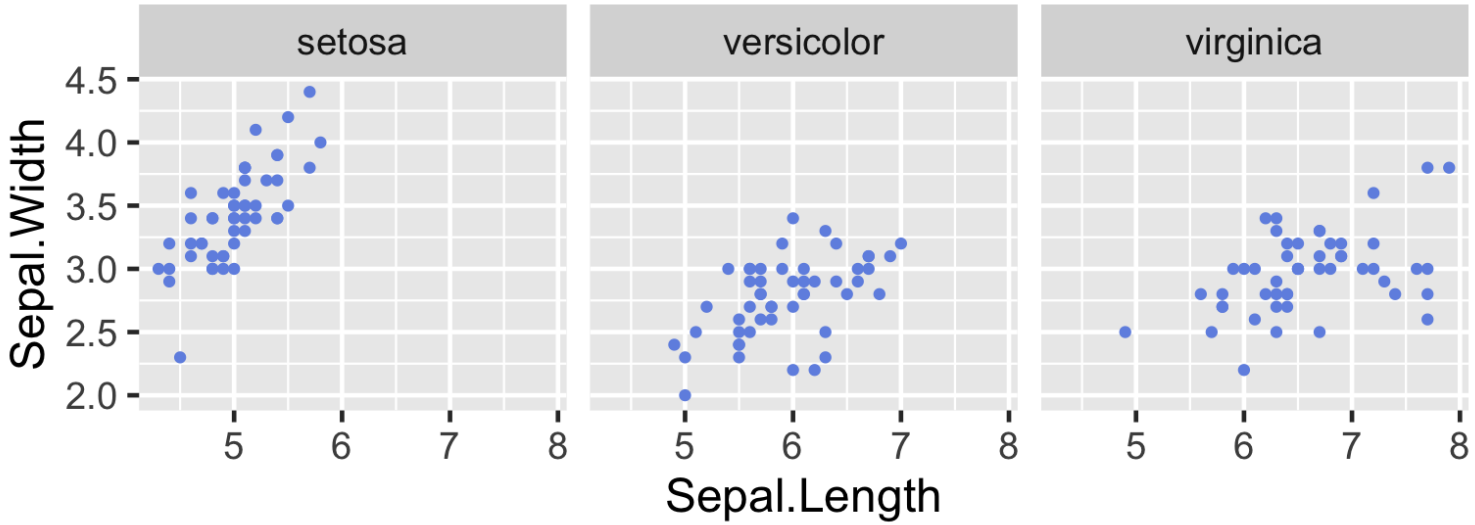
5.5 **Combining Continuous and Categorical Data**

6. Investingating Multivariate Continuous Data

7. Studying Multivariate Categorical Data

But first… FACETS

```
mycol = "#7192E3"
library(tidyverse)
ggplot(iris, aes(Sepal.Length, Sepal.Width)) +
  geom_point(color = mycol) +
  facet_wrap(~Species) +
  theme_grey(18)
```

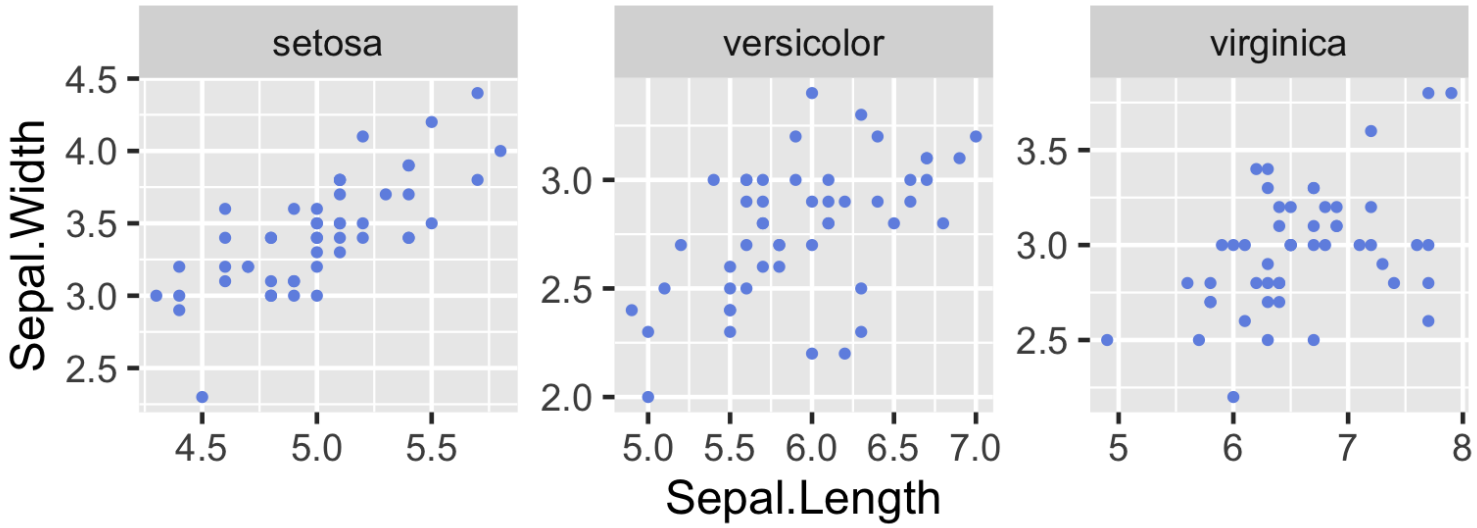


also called small multiples, panels, trellis, lattice

Main use: consistent scales – the default – for comparison

Less useful

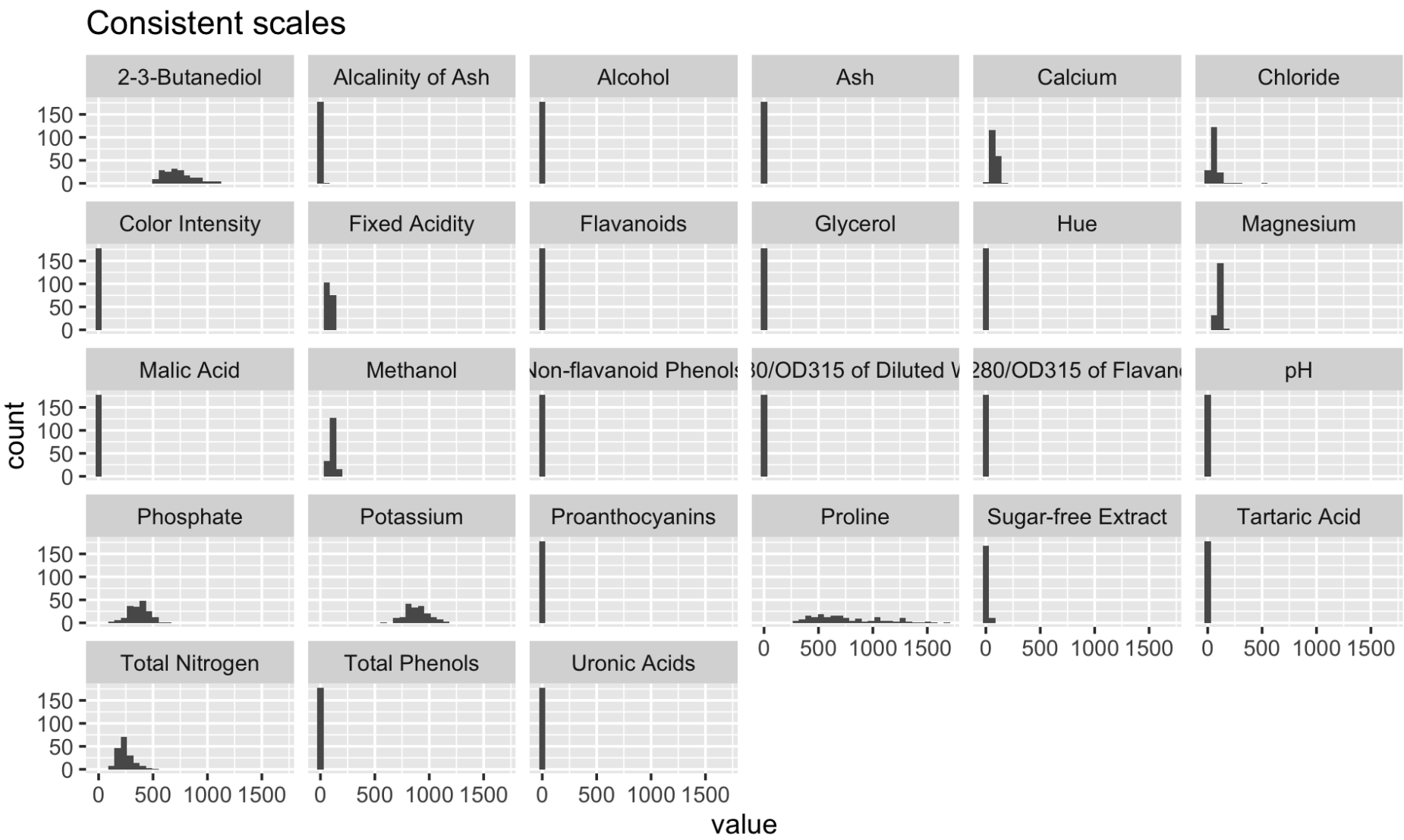
```
ggplot(iris, aes(Sepal.Length, Sepal.Width)) +
  geom_point(color = mycol) +
  facet_wrap(~Species, scales = "free") +
  theme_grey(18)
```



Faceting on variable rather than factor level

one panel for each numerical variable

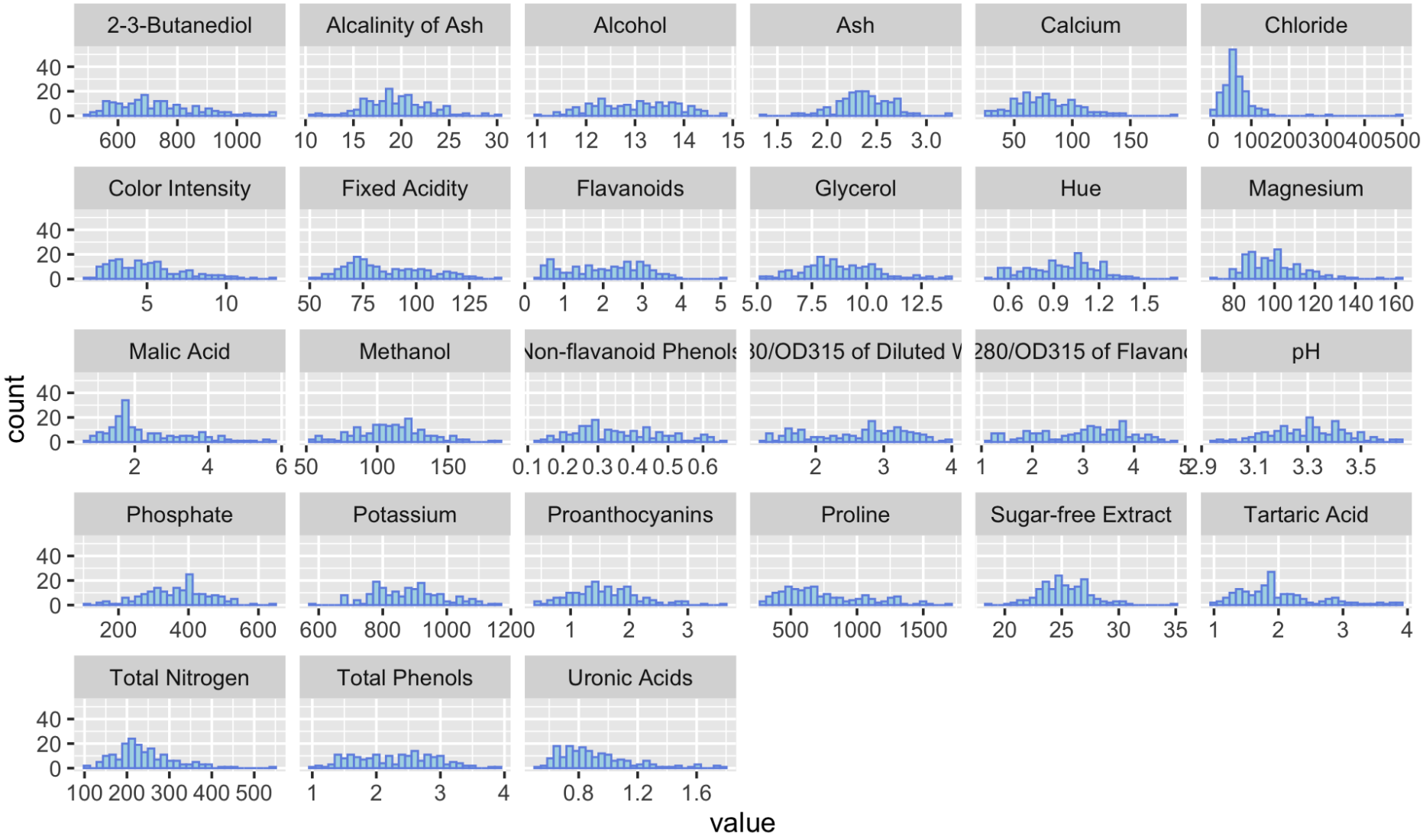
```
library(pgmm)
data(wine)
tidywine <- wine |>
  pivot_longer(cols = -Type, names_to = "variable", values_to = "value")
tidywine |>
  ggplot(aes(value)) +
  geom_histogram() +
  facet_wrap(~variable) +
  ggtitle("Consistent scales") +
  theme_grey(14)
```


Consistent scales
2-3-Butanediol
Alcalinity of Ash
Alcohol
Ash
Calcium
Chloride
Color Intensity
Fixed Acidity
Flavanoids
Glycerol
Hue
Magnesium
Malic Acid
Methanol
Non-flavanoid Phenols
280/OD315 of Diluted W
280/OD315 of Flavan
pH
Phosphate
Potassium
Proanthocyanins
Proline
Sugar-free Extract
Tartaric Acid
Total Nitrogen
Total Phenols
Uronic Acids
count
value
0
50
100
150
0
500
1000
1500

The exception that proves the rule

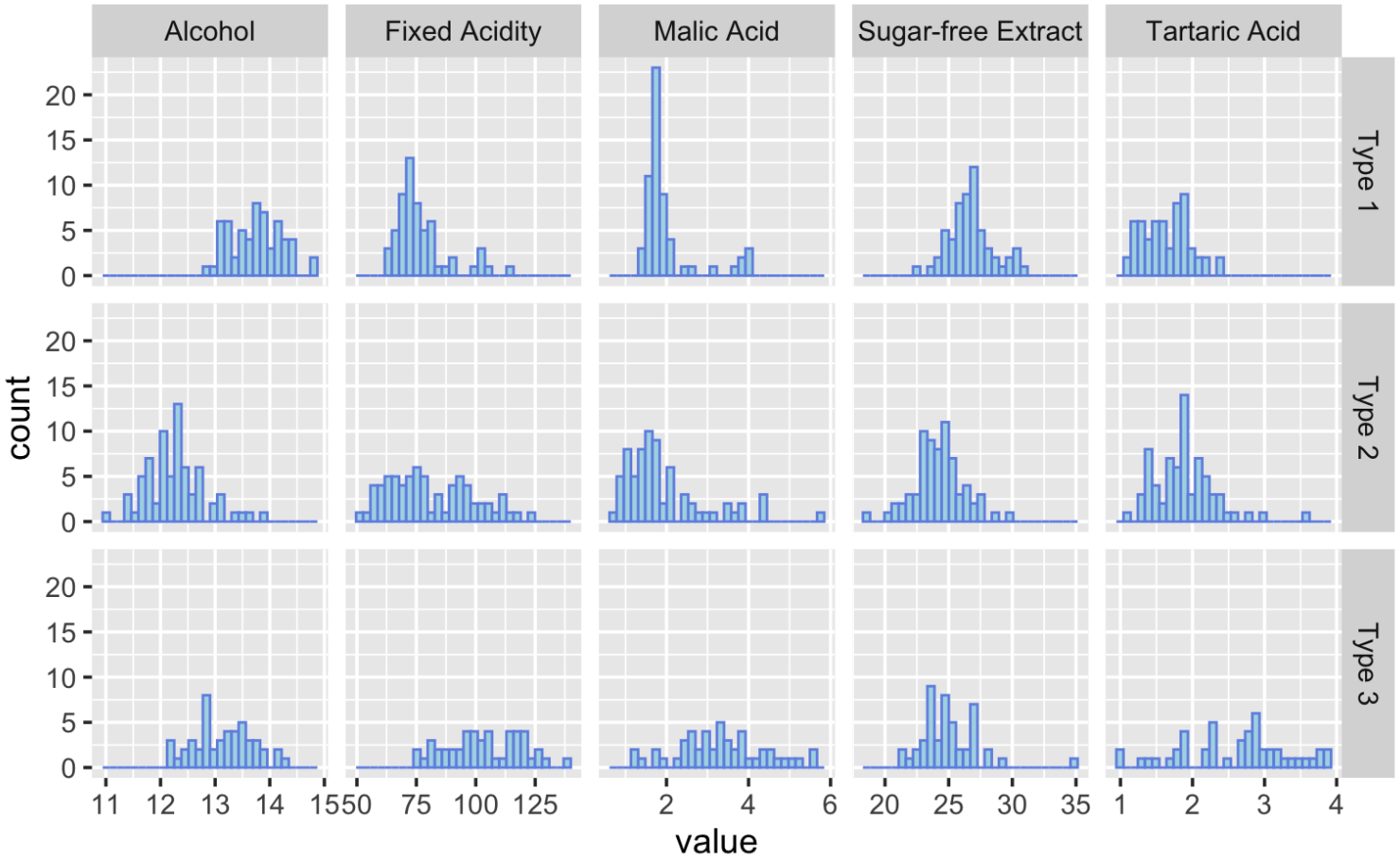
In this case, `scales = "free_x"` is a better option

```
tidywine |>
  ggplot(aes(value)) +
  geom_histogram(color = mycol, fill = "lightblue") +
  facet_wrap(~variable, scales = "free_x") +
  theme_grey(14)
```


2-3-Butanediol
Alcalinity of Ash
Alcohol
Ash
Calcium
Chloride
Color Intensity
Fixed Acidity
Flavanoids
Glycerol
Hue
Magnesium
Malic Acid
Methanol
Non-flavanoid Phenols
OD280/OD315 of Diluted W
OD280/OD315 of Flavan
pH
Phosphate
Potassium
Proanthocyanins
Proline
Sugar-free Extract
Tartaric Acid
Total Nitrogen
Total Phenols
Uronic Acids
count
value

Use `facet_grid` to add an additional variable

```
wine |>
  mutate(Type = paste("Type", Type)) |>
  select(1:6) |>
  pivot_longer(cols = -Type, names_to = "variable", values_to = "value") |>
  ggplot(aes(value)) +
  geom_histogram(color = mycol, fill = "lightblue") +
  facet_grid(Type ~ variable, scales = "free_x") +
  theme_grey(14)
```



```
library(GDAdata)
str(SpeedSki)
```

```
## 'data.frame':    91 obs. of  10 variables:
##  $ Rank      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Bib       : int  61 59 66 57 69 75 67 58 62 56 ...
##  $ FIS.Code  : int  7039 7078 190130 7178 510089 7204 7053 7170 7230 7055 ...
##  $ Name      : Factor w/ 91 levels "ABRAHAMSSON Mats",..: 64 63 56 83 54 13 66 14 19 43 ...
##  $ Year      : int  1979 1987 1985 1979 1970 1993 1975 1991 1980 1982 ...
##  $ Nation    : Factor w/ 14 levels "AUT","BEL","CAN",..: 7 7 5 1 12 5 13 5 4 13 ...
##  $ Speed     : num  212 210 210 210 209 ...
##  $ Sex       : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Event     : Factor w/ 3 levels "Speed Downhill",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ no.of.runs: int  4 4 4 4 4 4 4 4 4 4 ...
```

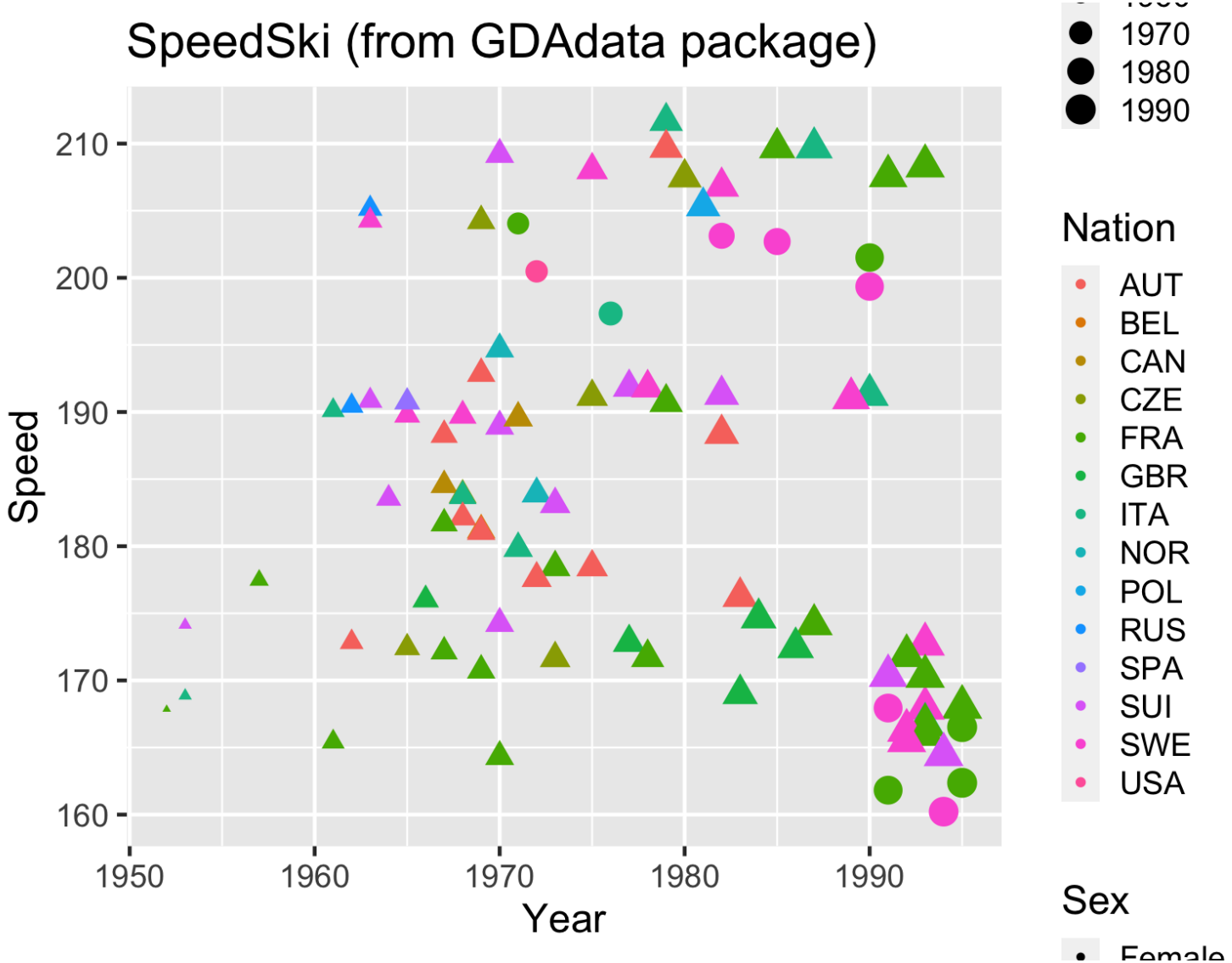
Mapping options

Continuous: x-axis, y-axis, color (not so great), size (not so great)

Categorical: color, facets (rows, columns), shape (maybe)

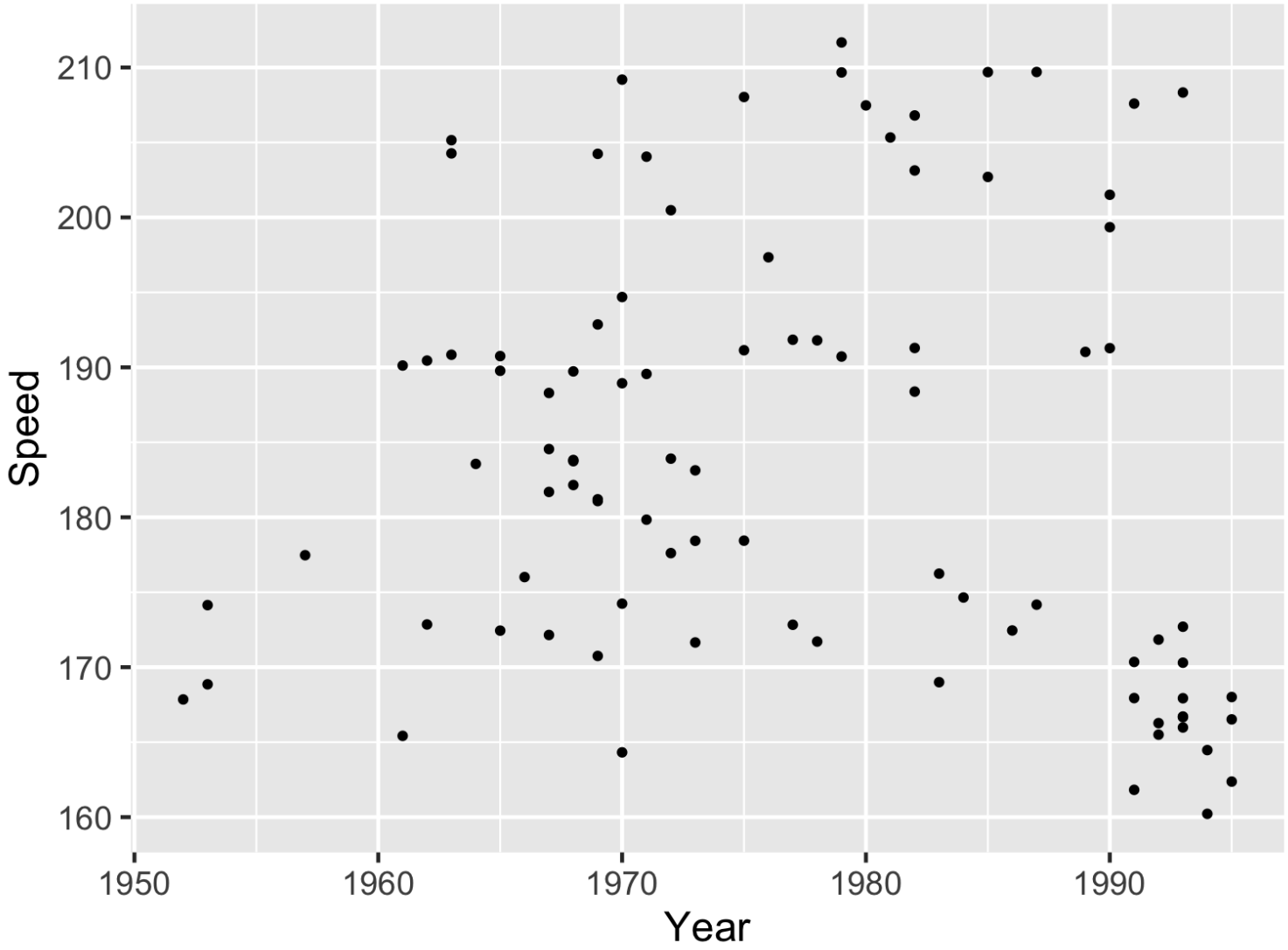
Don’t over do it

```
ggplot(SpeedSki, aes(Year, Speed, size = Year, shape = Sex, color = Nation)) +
  geom_point() +
  ggtitle("SpeedSki (from GDAdata package)") +
  theme_grey(18)
```


SpeedSki (from GDAdata package)
Speed
210
200
190
180
170
160
1950
1960
1970
1980
1990
Year
1970
1980
1990
Nation
AUT
BEL
CAN
CZE
FRA
GBR
ITA
NOR
POL
RUS
SPA
SUI
SWE
USA
Sex
Female

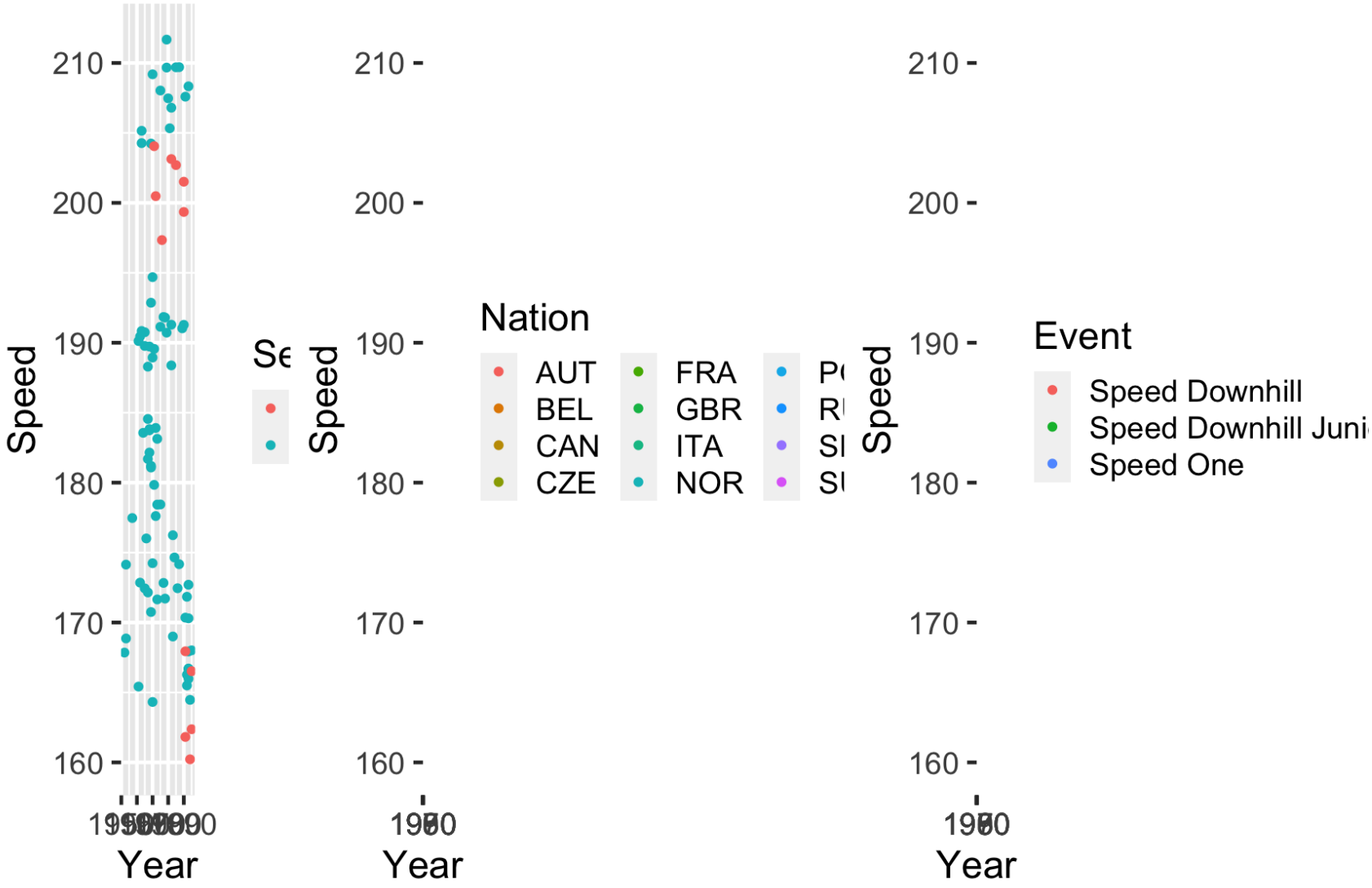
Start simple

```
ggplot(SpeedSki, aes(Year, Speed)) +
  geom_point() +
  theme_grey(18) +
  theme(legend.position = "bottom")
```


Speed
210
200
190
180
170
160
1950
1960
1970
1980
1990
Year

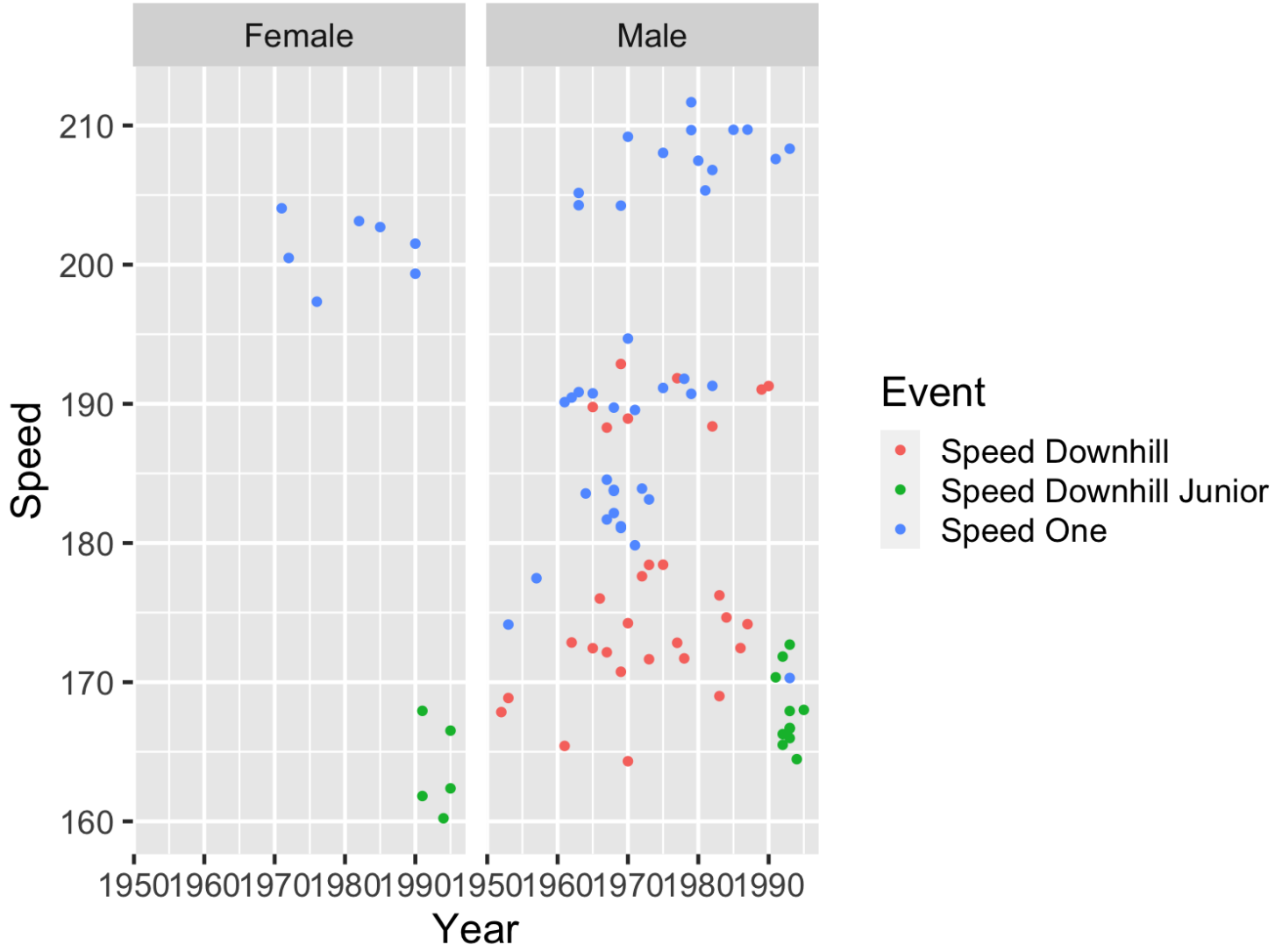
Add one variable at a time

```
g <- ggplot(SpeedSki, aes(Year, Speed)) +
  theme(legend.position = "bottom", legend.direction = "vertical") +
  theme_grey(18)

gSex <- g +
  geom_point(aes(color = Sex))

gNation <- g +
  geom_point(aes(color = Nation)) +
  guides(color=guide_legend(ncol=4))

gEvent <- g +
  geom_point(aes(color = Event))

gridExtra::grid.arrange(gSex, gNation, gEvent, nrow = 1)
```


Speed
210
200
190
180
170
160
Year
Se
Nation
AUT
BEL
CAN
CZE
FRA
GBR
ITA
NOR
Event
Speed Downhill
Speed Downhill Juni
Speed One

Facet

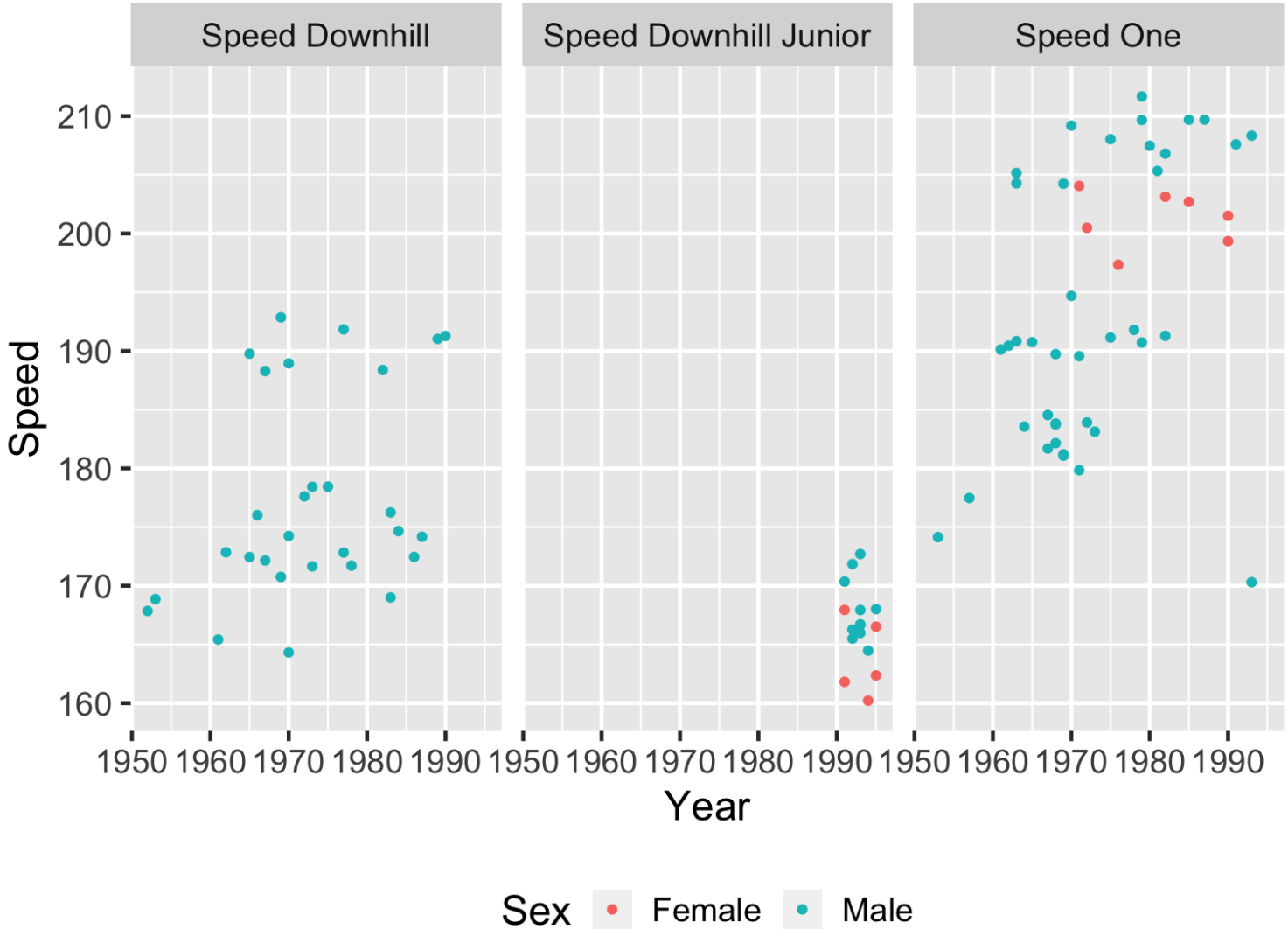
```
ggplot(SpeedSki, aes(Year, Speed)) +
  geom_point(aes(color = Event)) +
  facet_wrap(~Sex) +
  theme_grey(16) +
  theme(legend.position = "bottom") +
  theme_grey(18)
```



Note the consistent scales

Try switching faceting and color

```
ggplot(SpeedSki, aes(Year, Speed, color = Sex)) +
  geom_point() +
  facet_wrap(~Event) +
  theme_grey(18) +
  theme(legend.position = "bottom")
```


Speed Downhill
Speed Downhill Junior
Speed One
Speed
210
200
190
180
170
160
1950 1960 1970 1980 1990
Year
Sex
Female
Male