Data cleaning and exploration in R

Nsubuga Emmnauel Reagan

2024-04-24

```
#install.packages(c("nycflights13", "gapminder", "Lahman"))
#library(nycflights13)
#library(gapminder)
#library(Lahman)
```

Data visua;isation with the Tidyverse

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

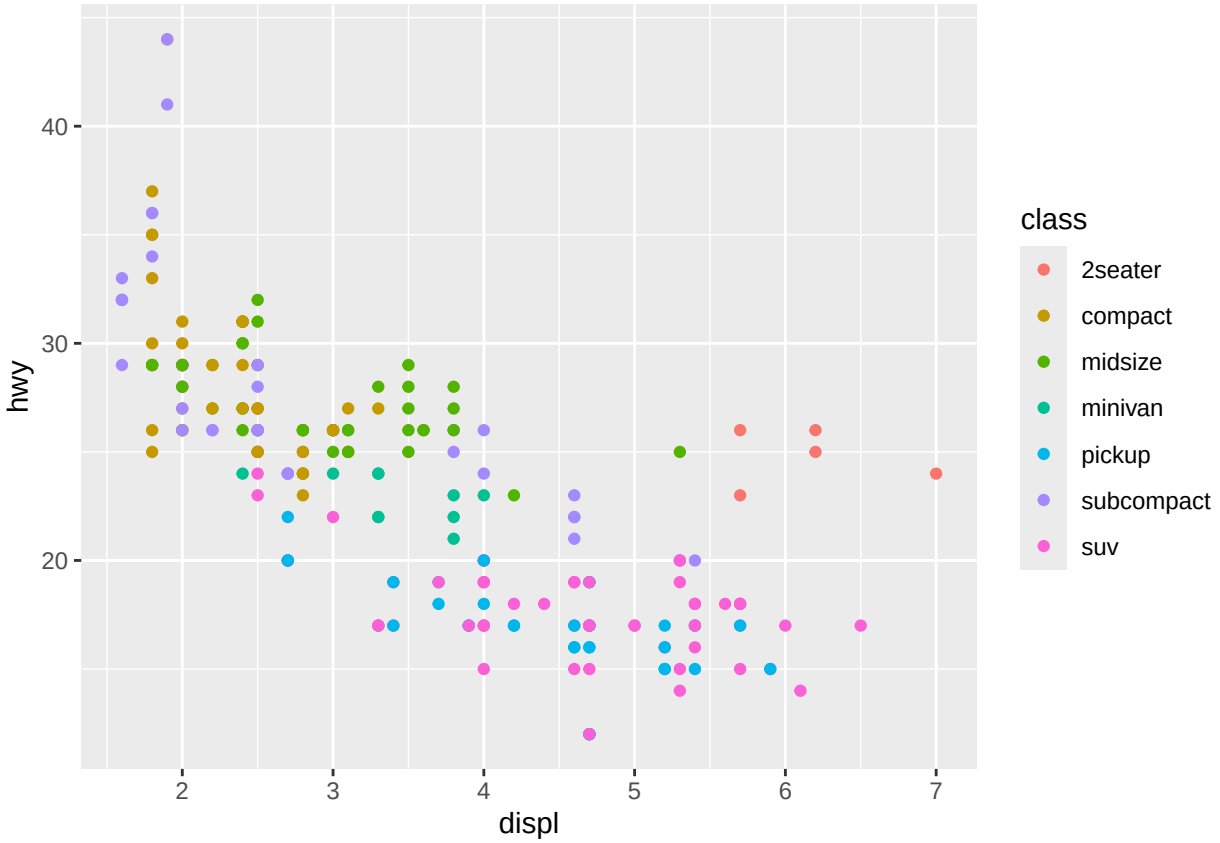
```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
## Warning: package 'stringr' was built under R version 4.3.3
```

```
## Warning: package 'lubridate' was built under R version 4.3.3
```

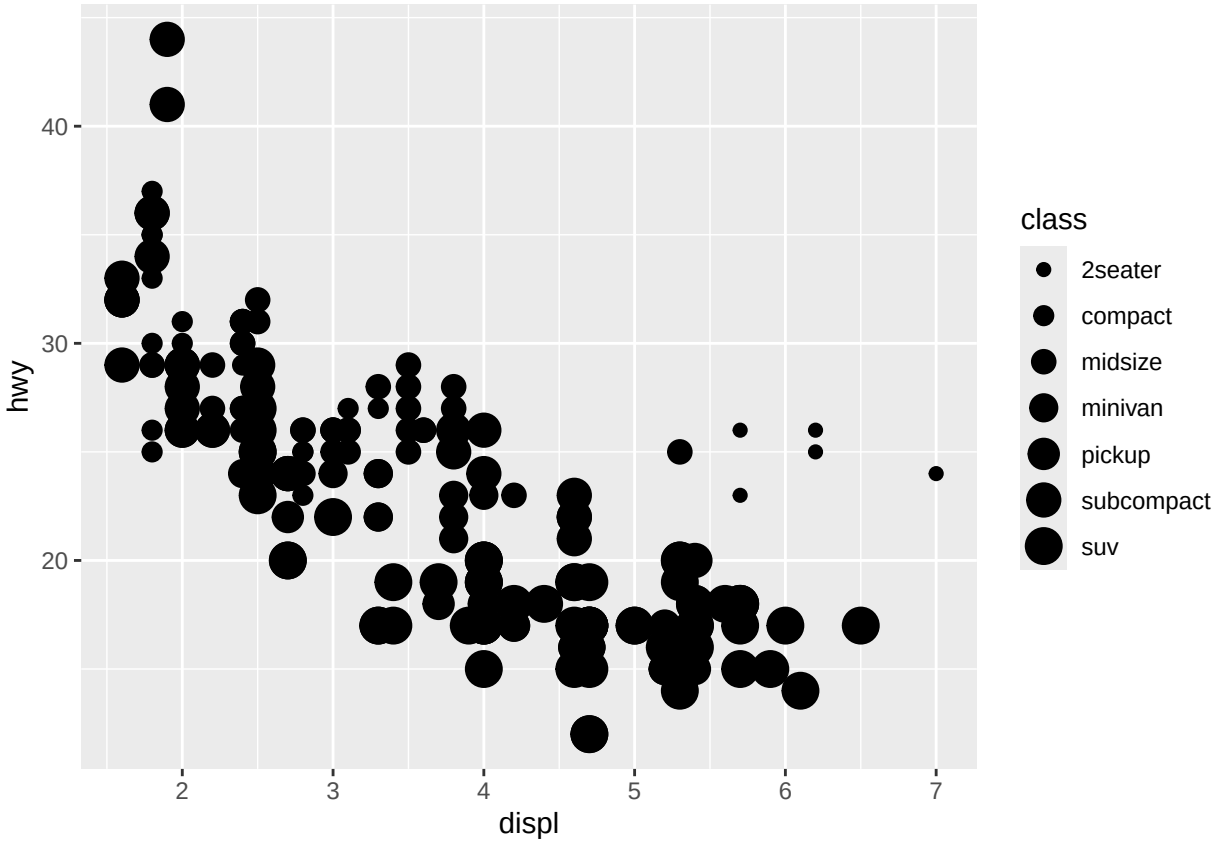
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, color = class))
```



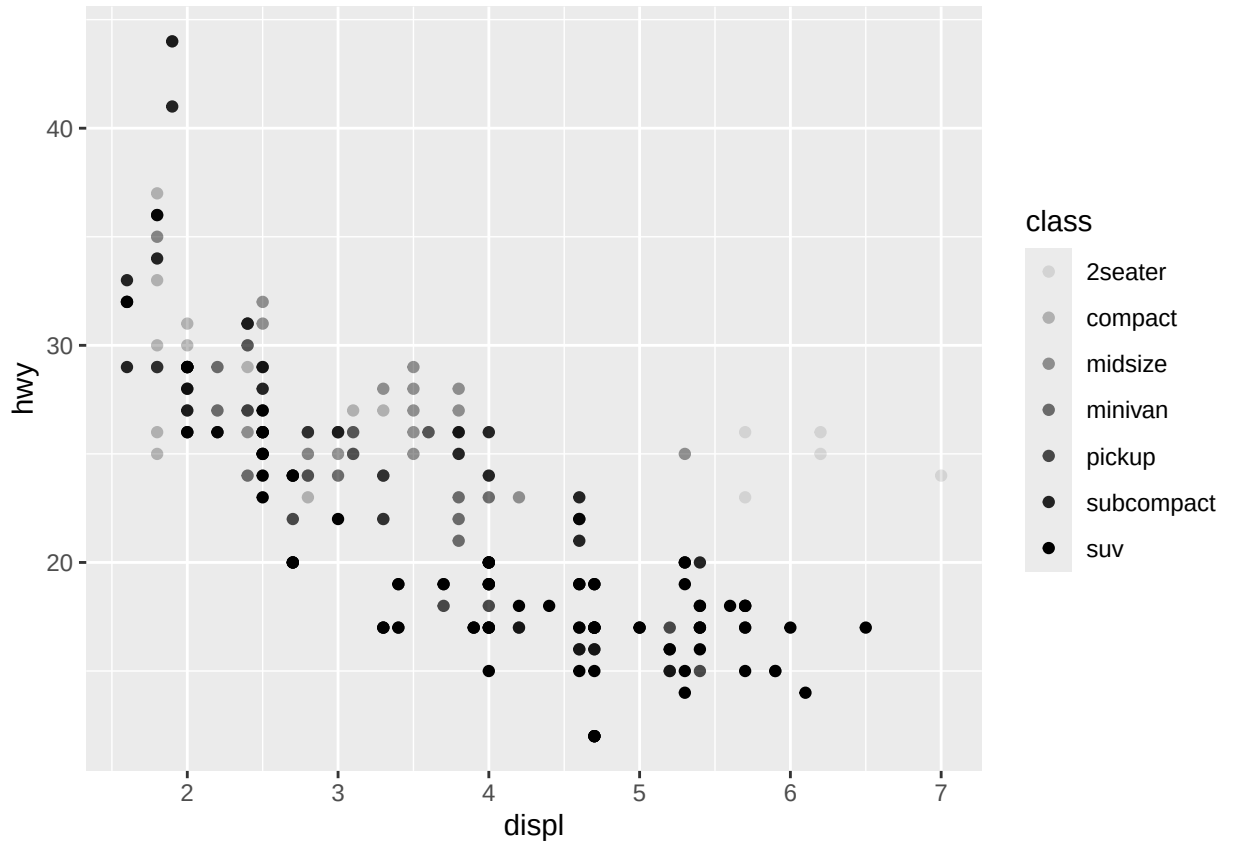

```
ggplot(data = mpg) +
 geom_point(mapping = aes(x = displ, y = hwy, size = class))
```

```
## Warning: Using size for a discrete variable is not advised.
```

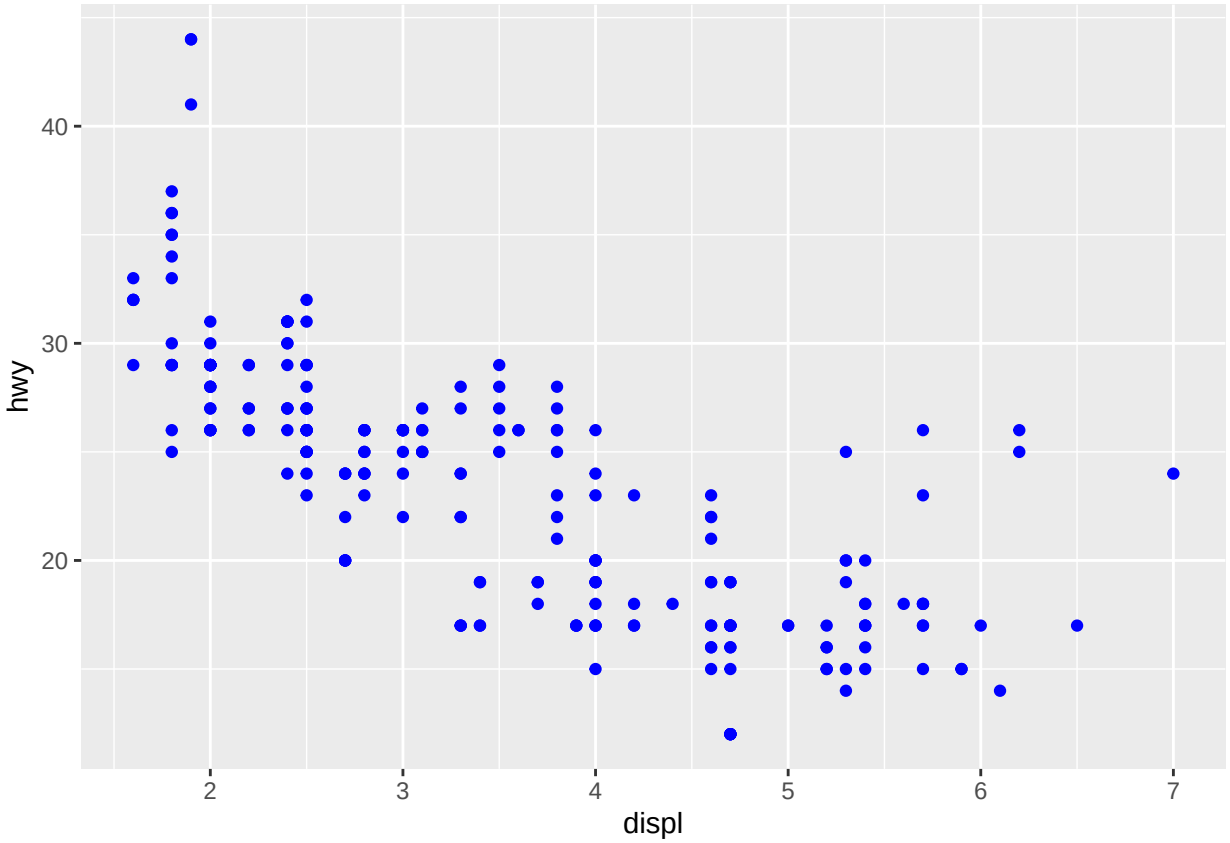



```
ggplot(data = mpg) +
geom_point(mapping = aes(x = displ, y = hwy, alpha = class))
```

```
## Warning: Using alpha for a discrete variable is not advised.
```

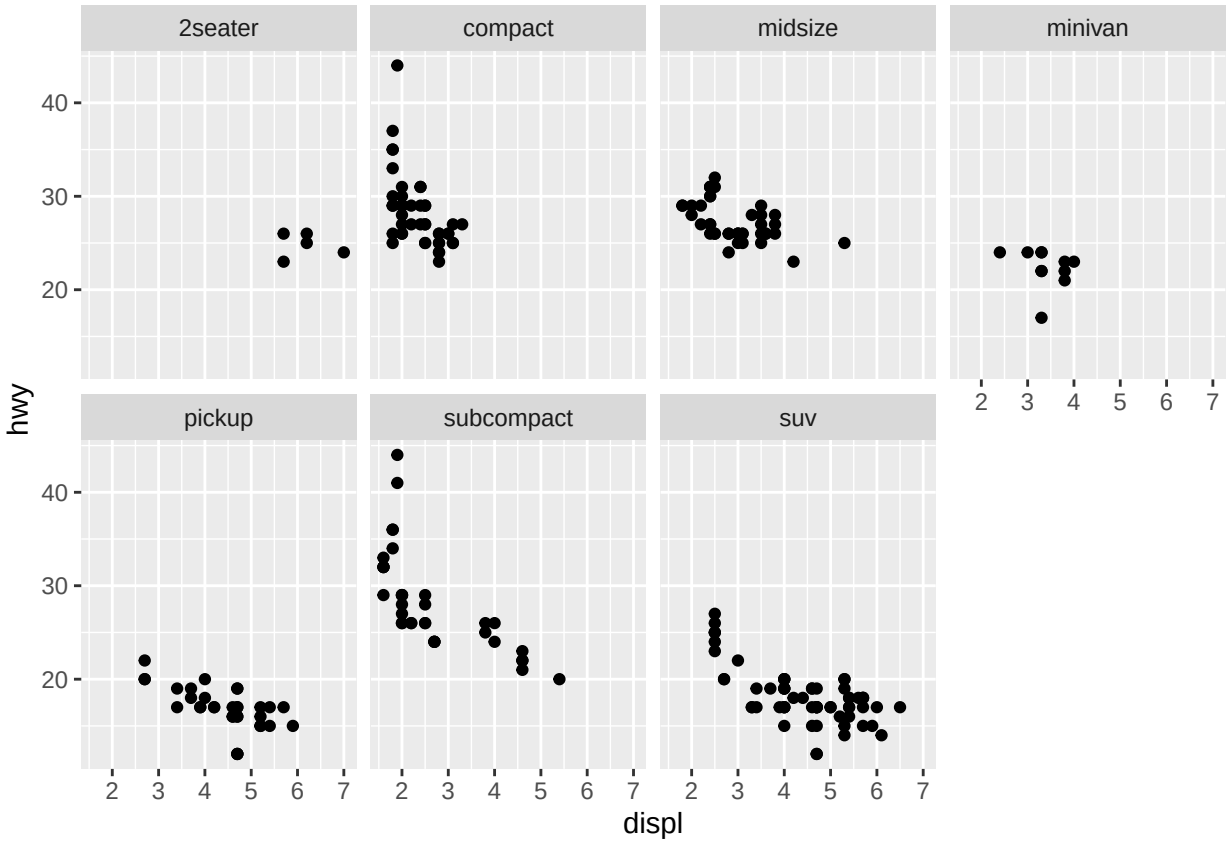



```
ggplot(data = mpg) +
 geom_point(mapping = aes(x = displ, y = hwy), color = "blue")
```

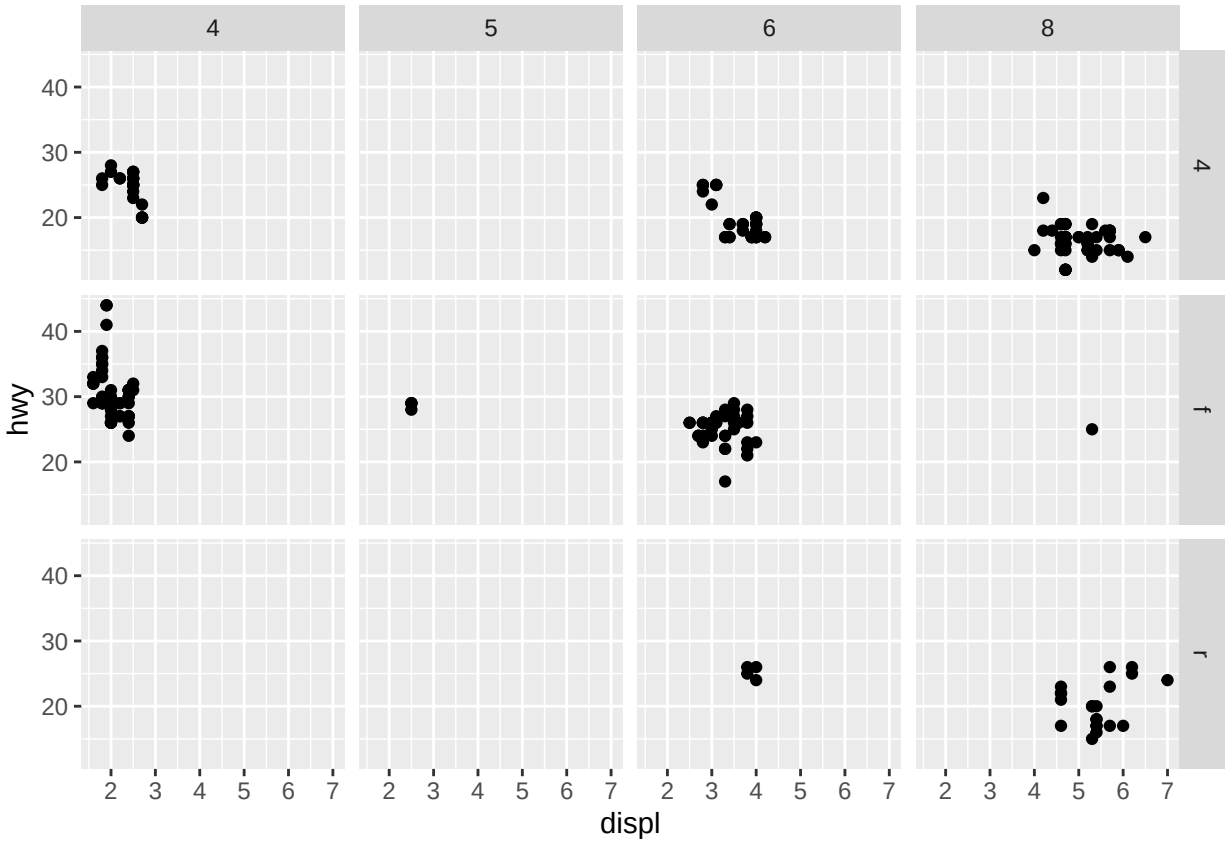


Facets

```
ggplot(data = mpg) +
 geom_point(mapping = aes(x = displ, y = hwy)) +
 facet_wrap(~ class, nrow = 2)
```

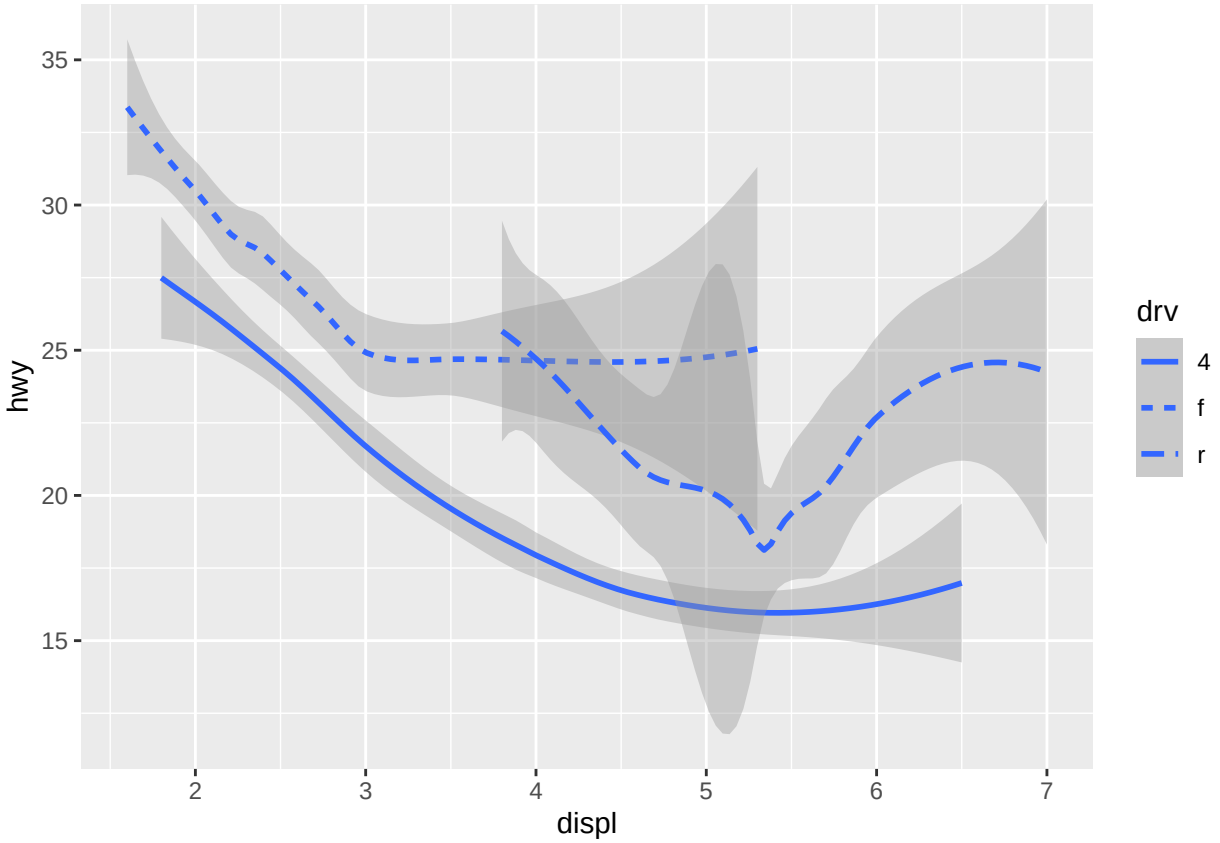



```
ggplot(data = mpg) +
 geom_point(mapping = aes(x = displ, y = hwy)) +
 facet_grid(drv ~ cyl)
```



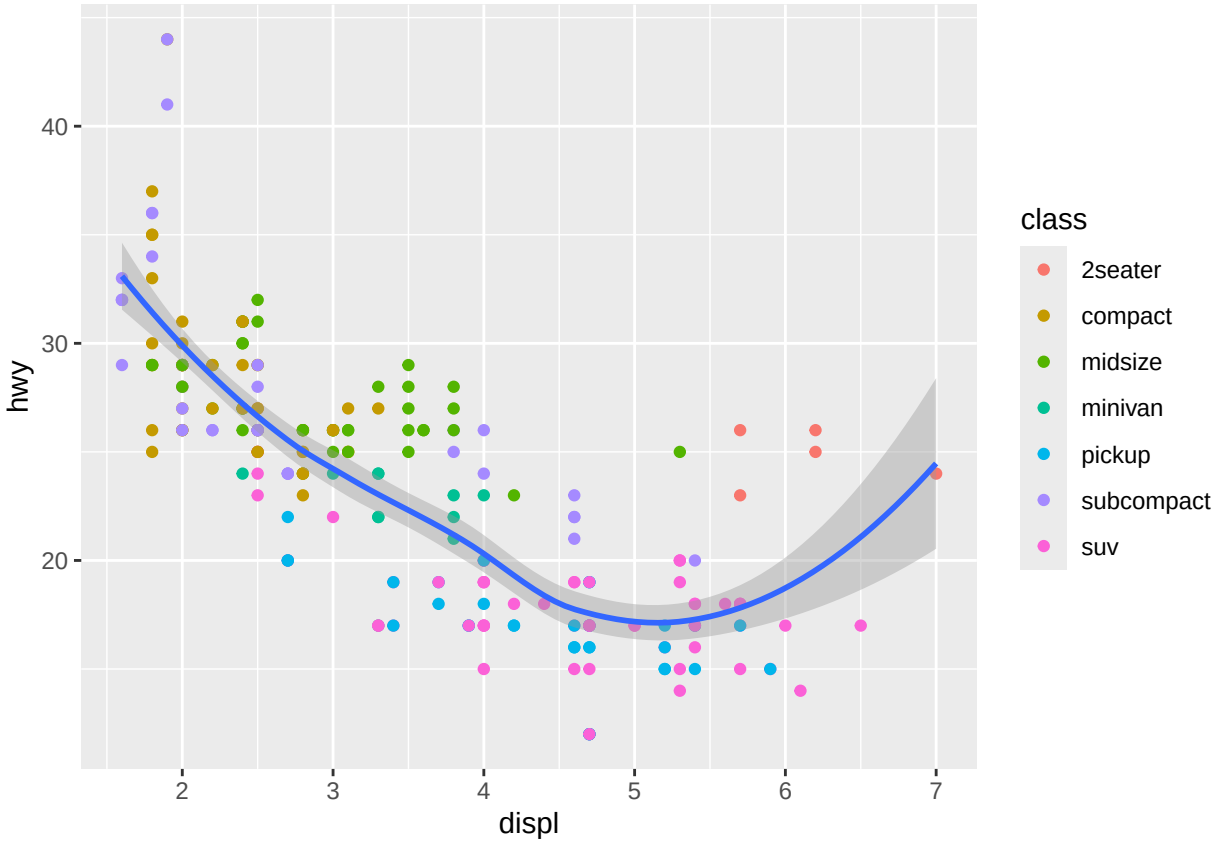

```
ggplot(data = mpg) +
geom_smooth(mapping = aes(x = displ, y = hwy, linetype = drv))
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



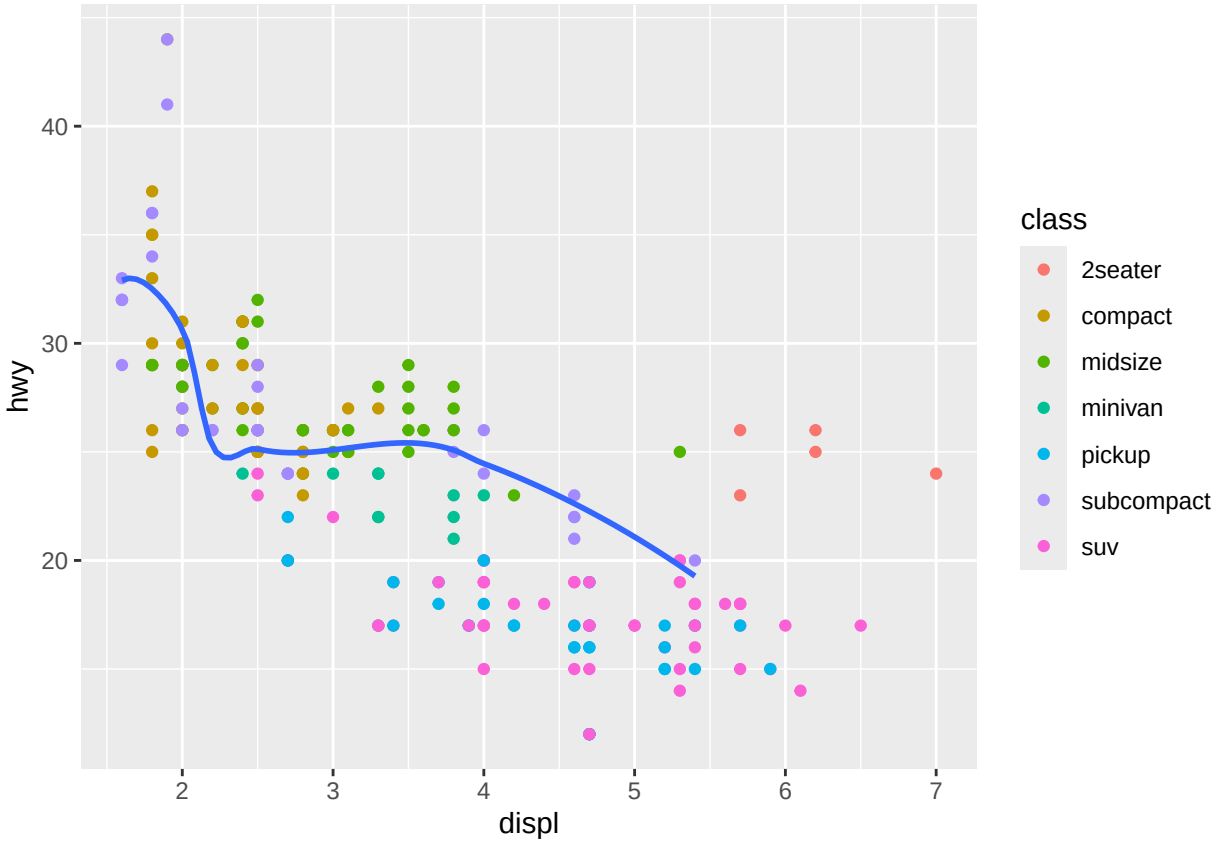

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
geom_point(mapping = aes(color = class)) +
geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

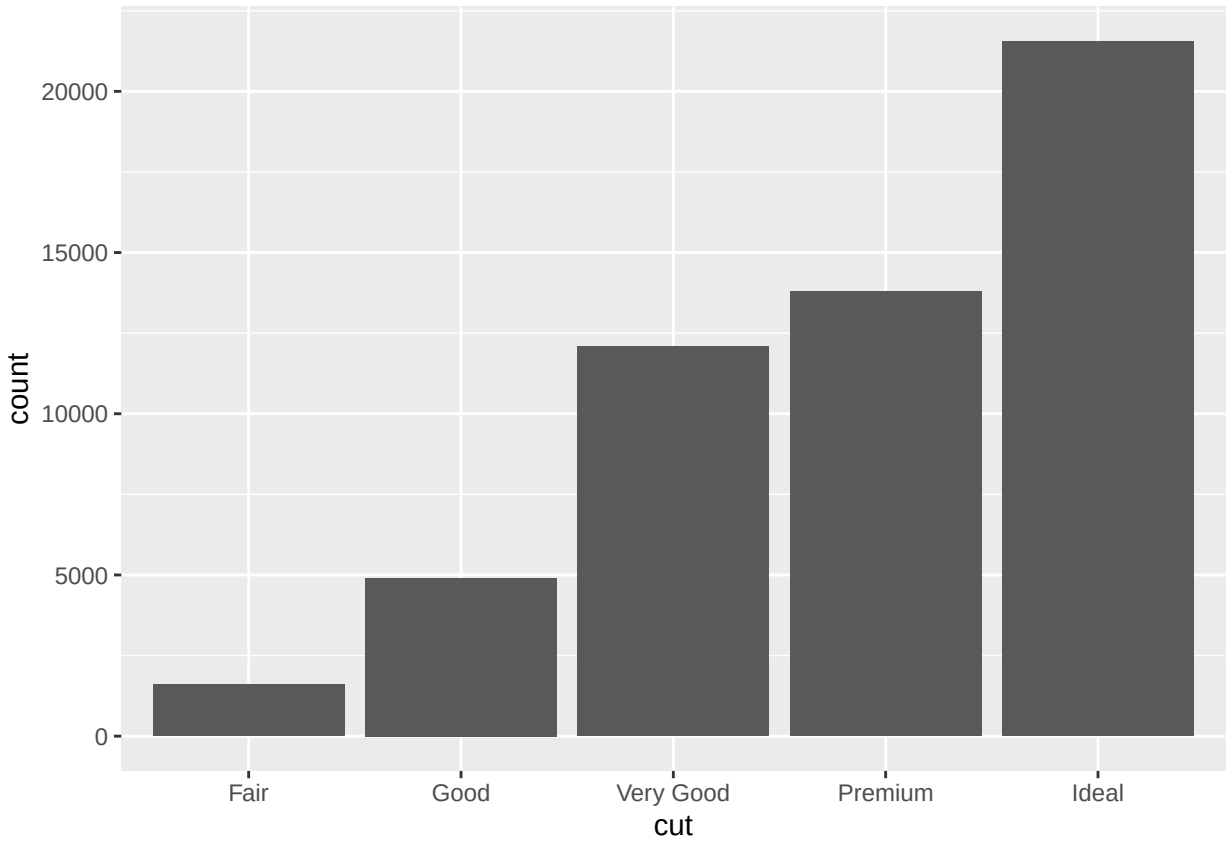



```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
geom_point(mapping = aes(color = class)) +
geom_smooth(
data = filter(mpg, class == "subcompact"),
se = FALSE
)
```

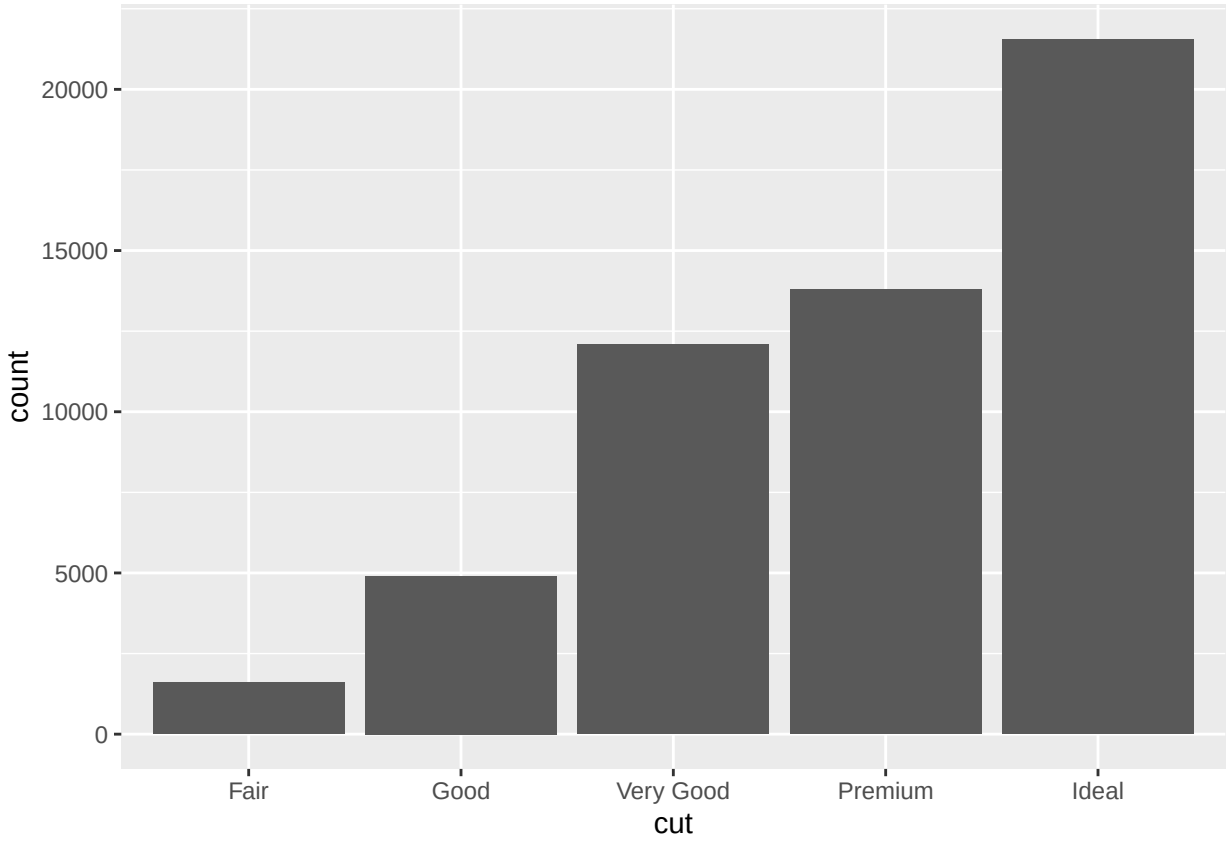
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




```
ggplot(data = diamonds) +
geom_bar(mapping = aes(x = cut))
```

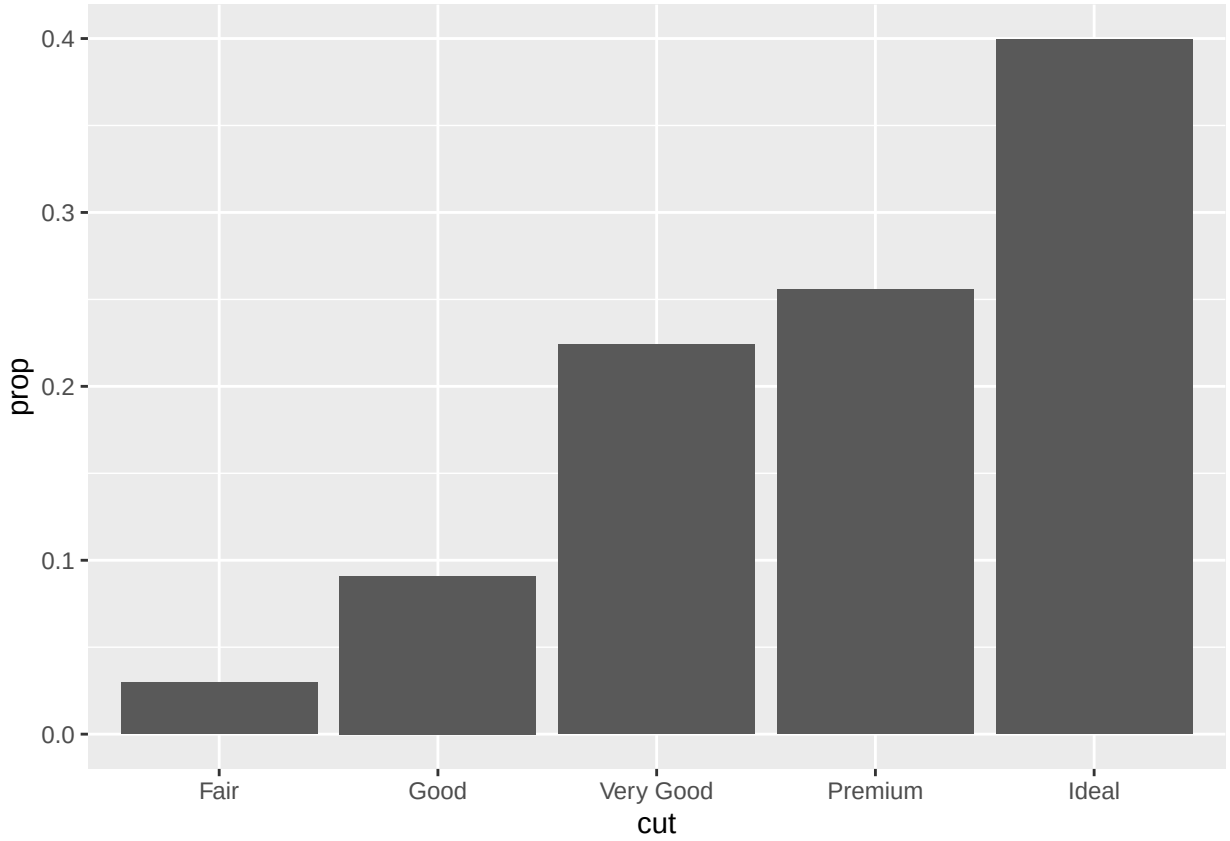



```
ggplot(data = diamonds) +
 stat_count(mapping = aes(x = cut))
```




```
ggplot(data = diamonds) +
geom_bar(
mapping = aes(x = cut, y = ..prop.., group = 1)
)
```

```
## Warning: The dot-dot notation (`..prop..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(prop)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

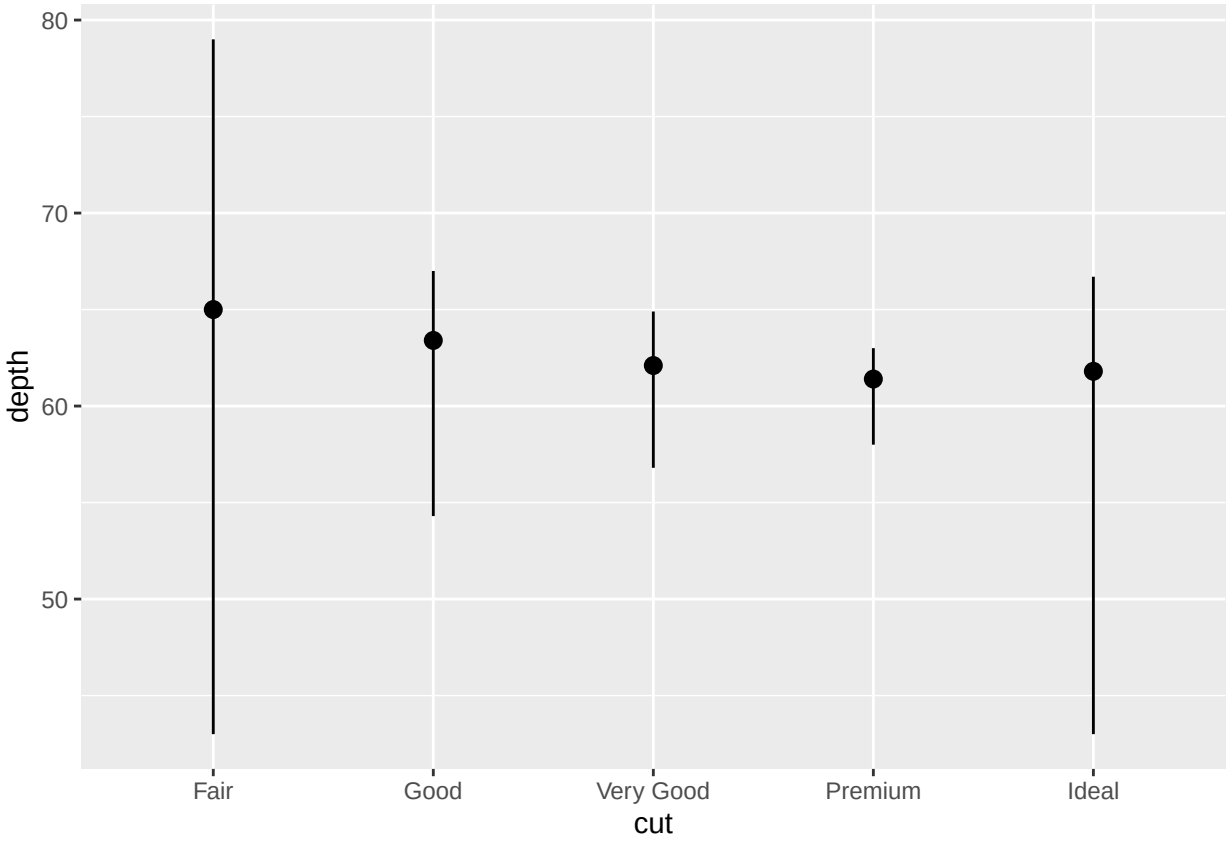



```
ggplot(data = diamonds) +
stat_summary(
mapping = aes(x = cut, y = depth),
fun.ymin = min,
fun.ymax = max,
fun.y = median
)
```

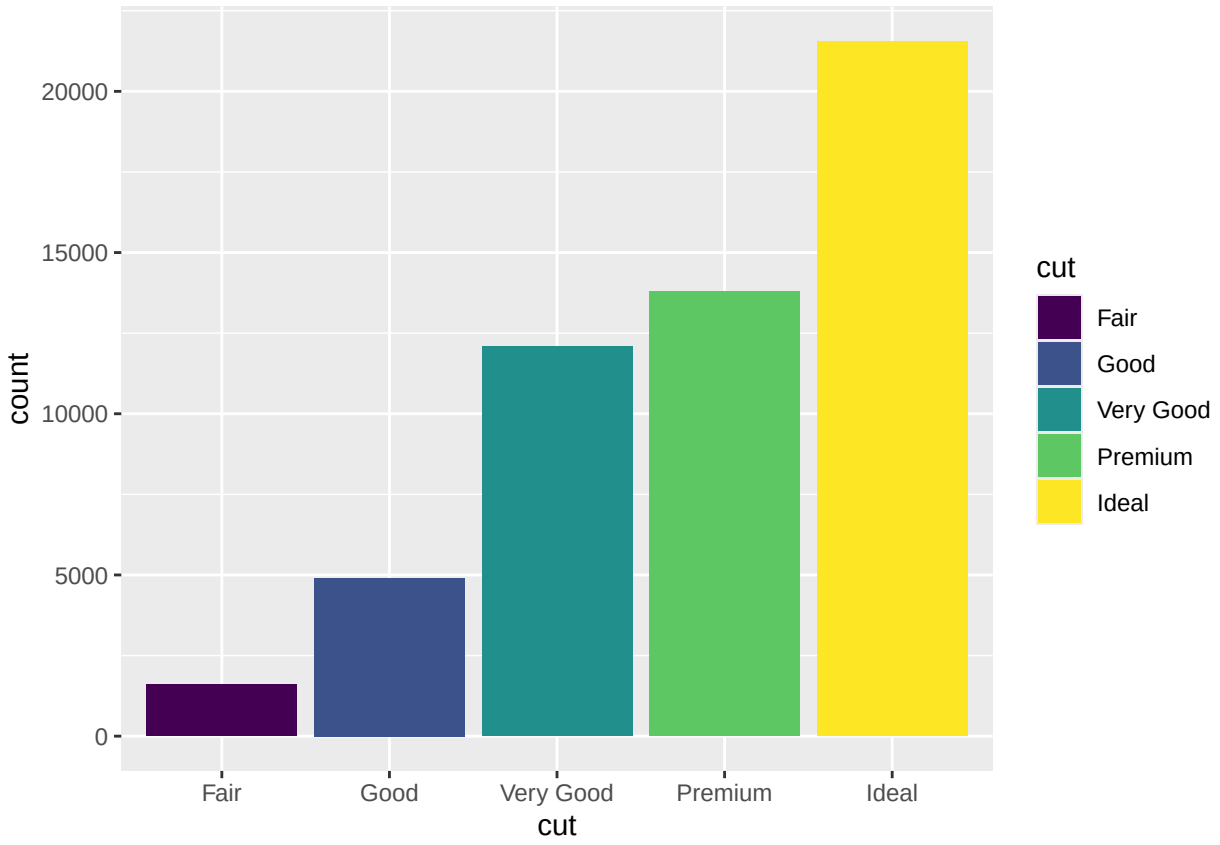
```
## Warning: The `fun.y` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
## i Please use the `fun` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The `fun.ymin` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
## i Please use the `fun.min` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

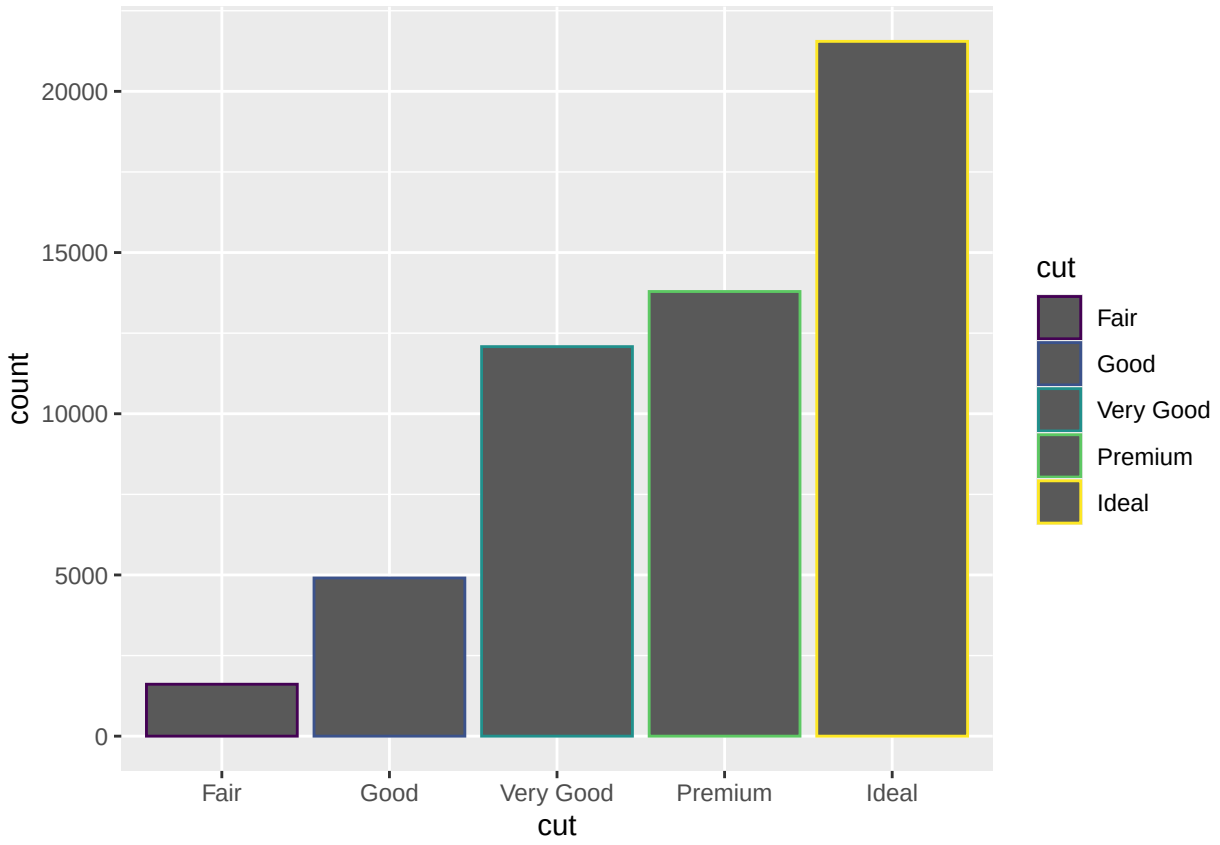
```
## Warning: The `fun.ymax` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
## i Please use the `fun.max` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



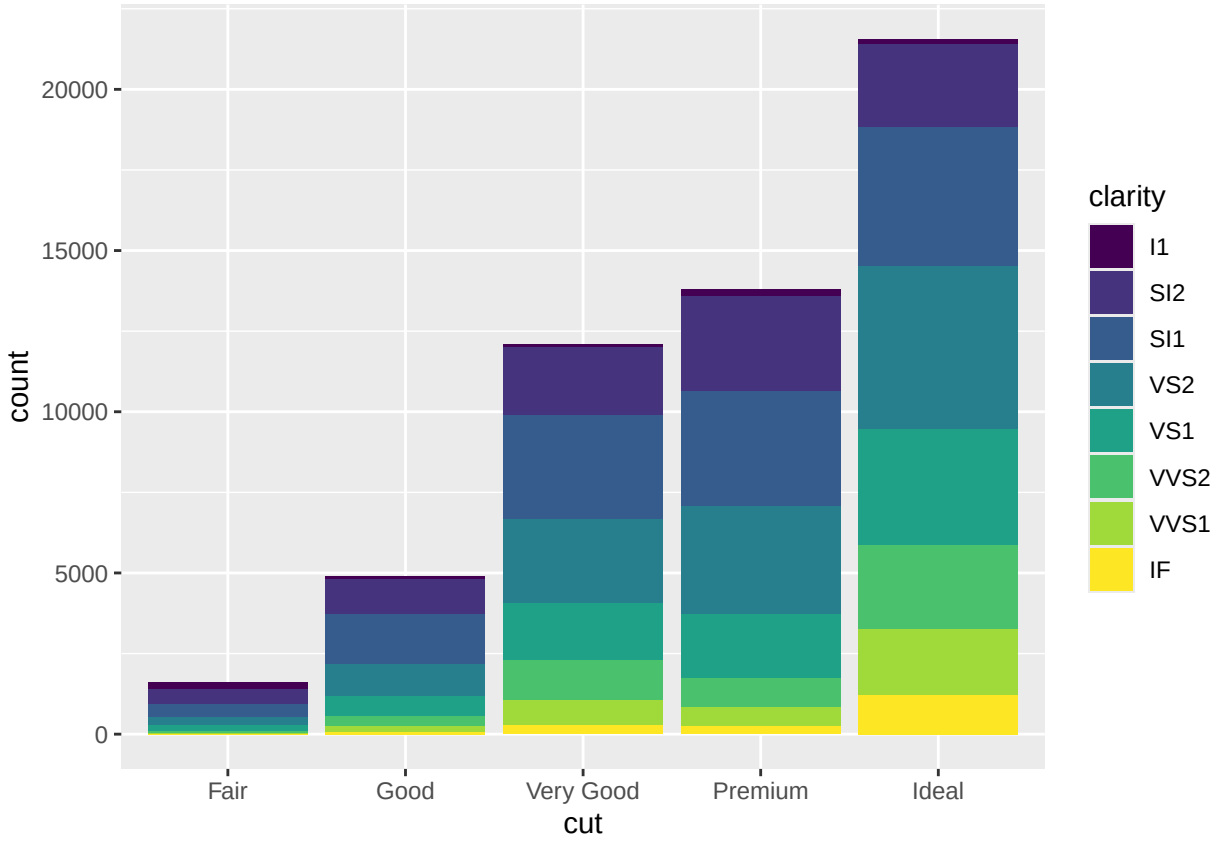

```
ggplot(data = diamonds) +
geom_bar(mapping = aes(x = cut, fill = cut))
```



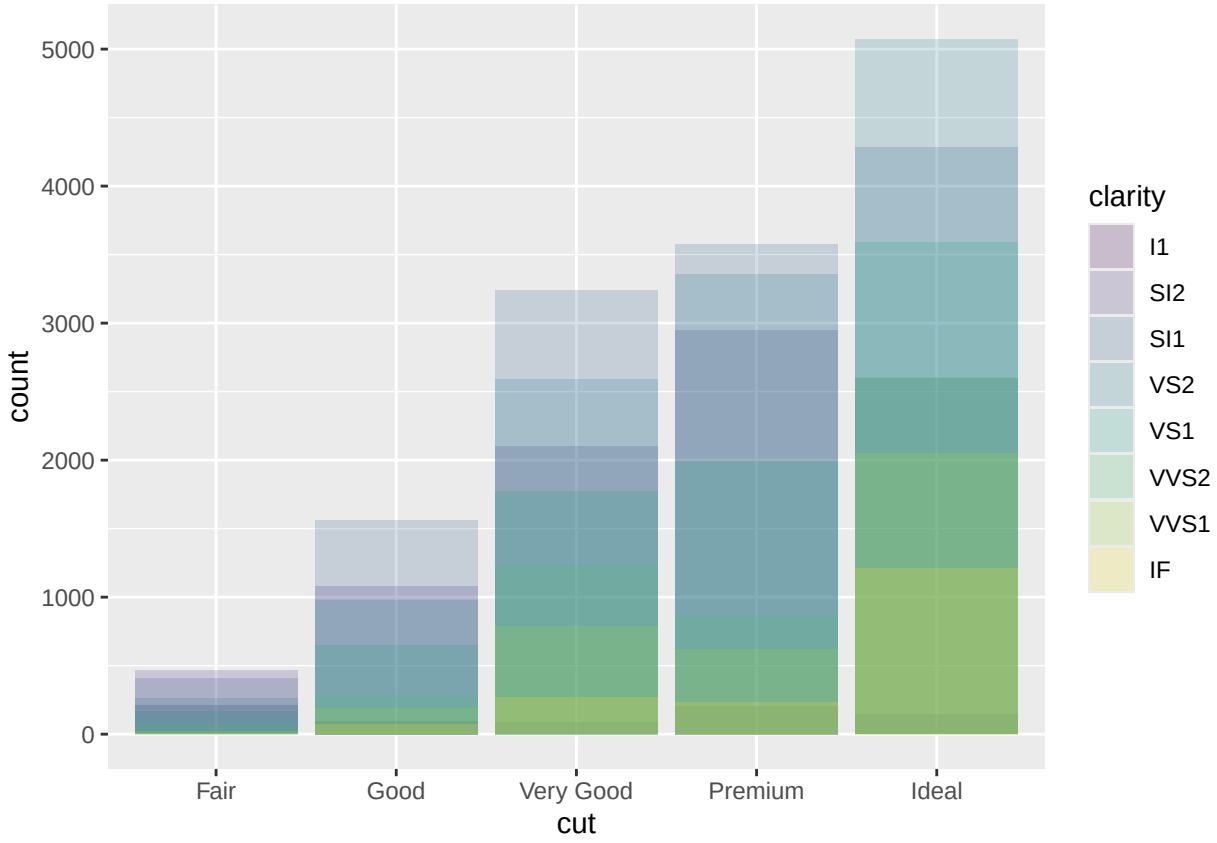

```
ggplot(data = diamonds) +
 geom_bar(mapping = aes(x = cut, color = cut))
```



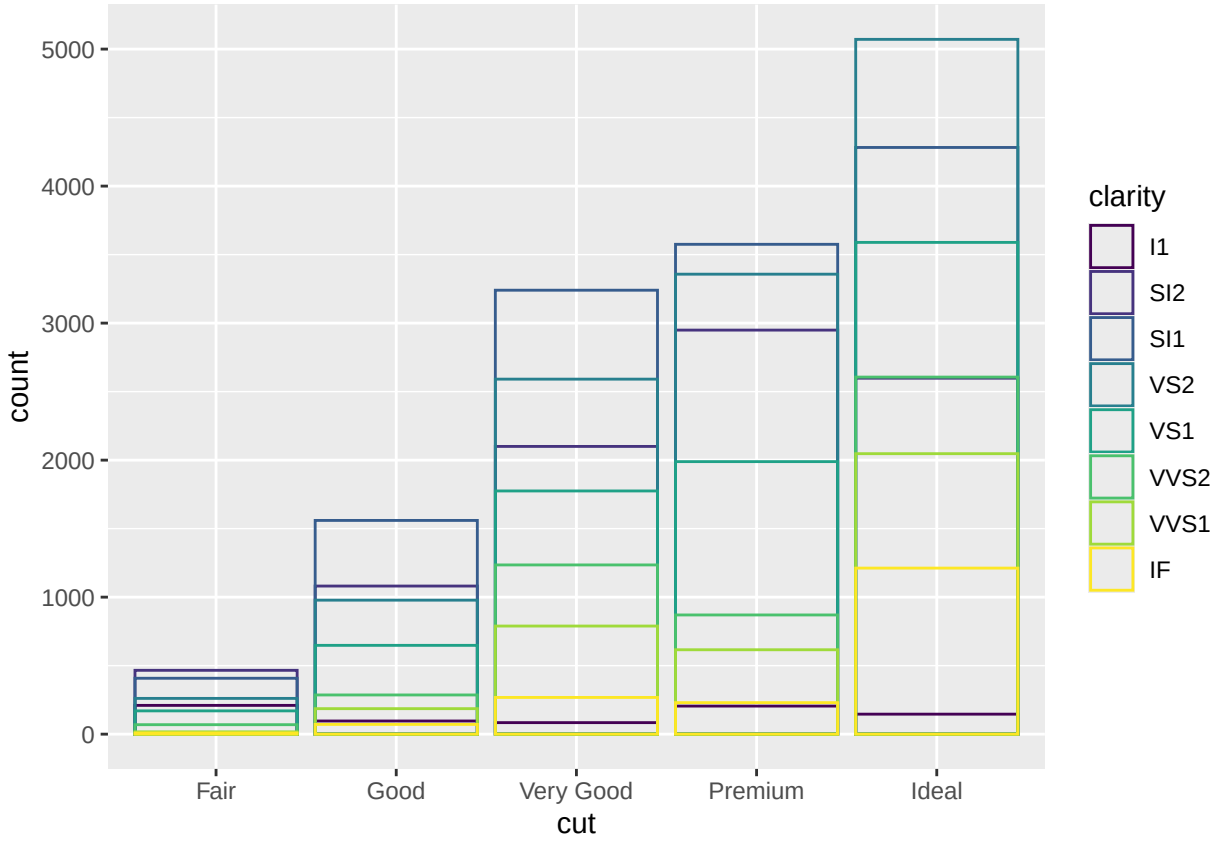

```
ggplot(data = diamonds) +
 geom_bar(mapping = aes(x = cut, fill = clarity))
```



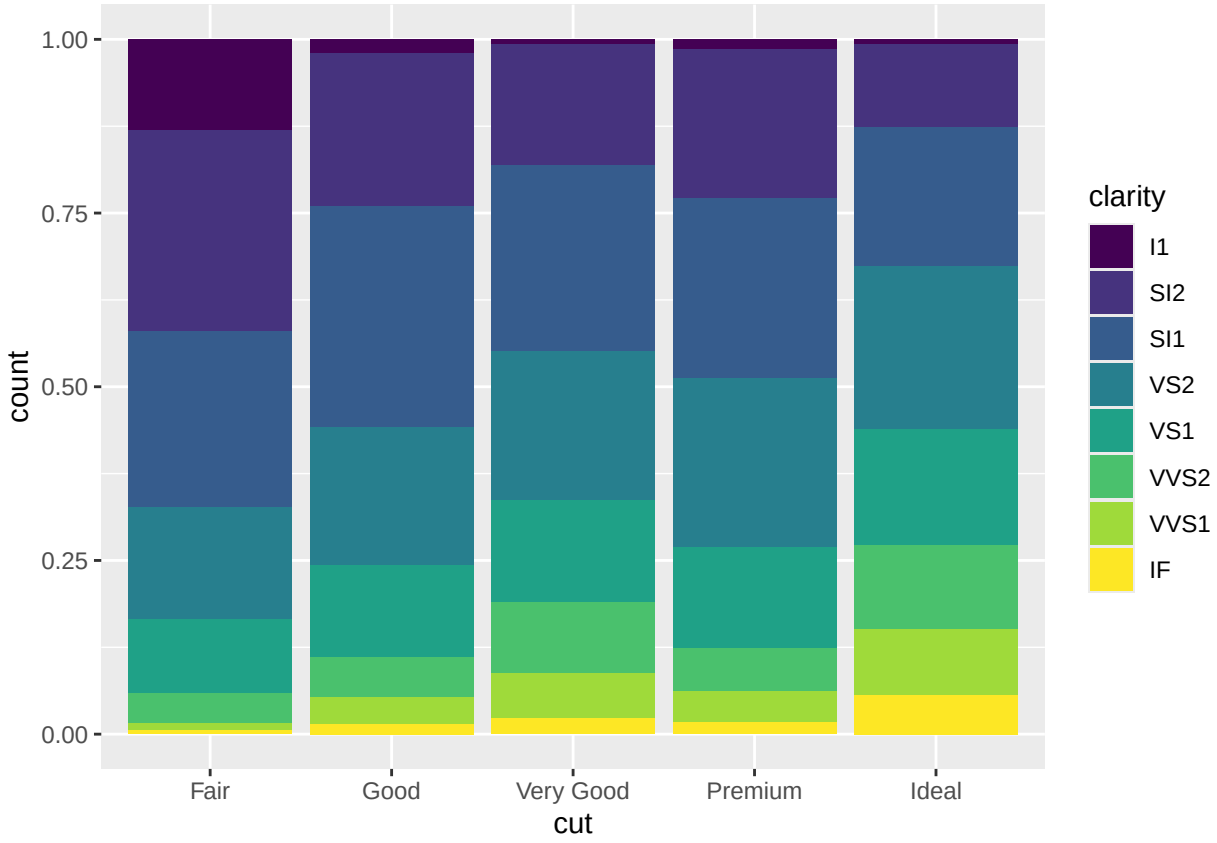

```
ggplot(
data = diamonds,
mapping = aes(x = cut, fill = clarity)
) +
geom_bar(alpha = 1/5, position = "identity")
```



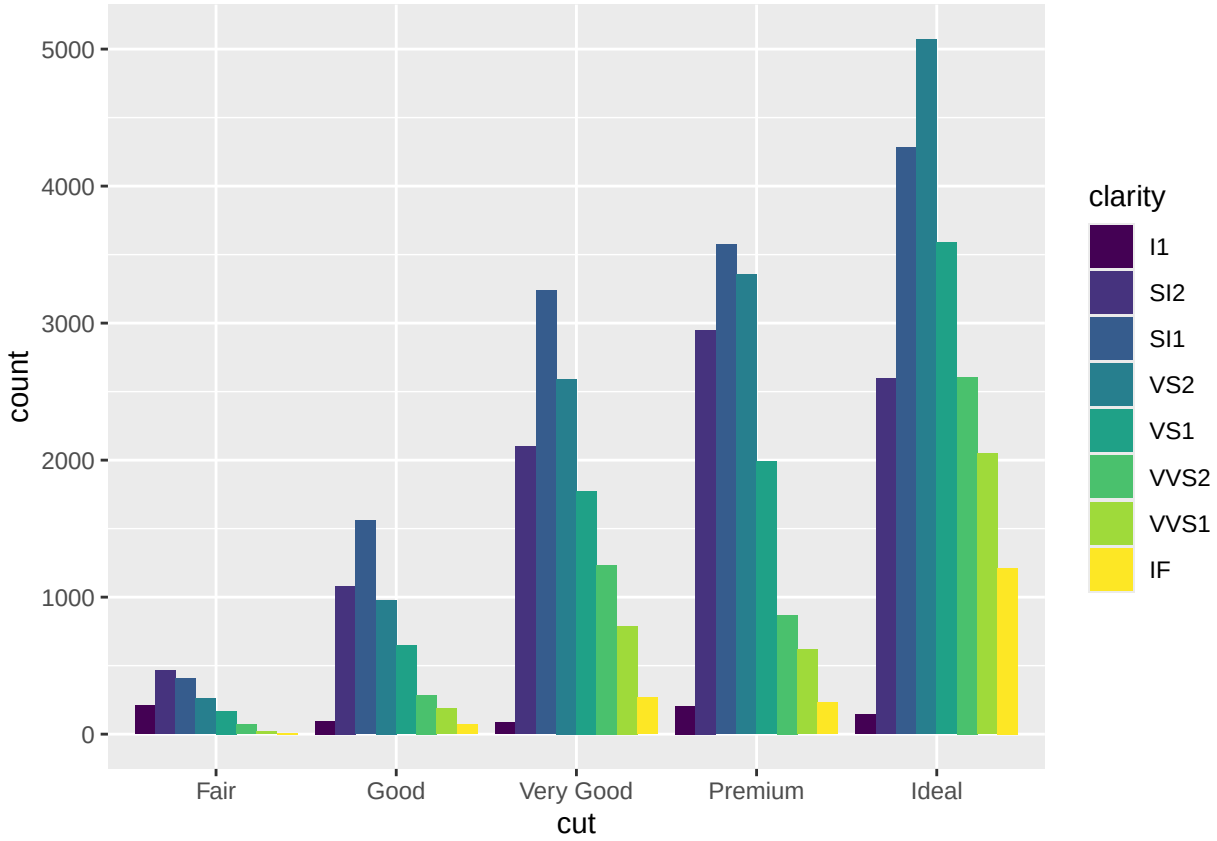

```
ggplot(
data = diamonds,
mapping = aes(x = cut, color = clarity)
) +
geom_bar(fill = NA, position = "identity")
```



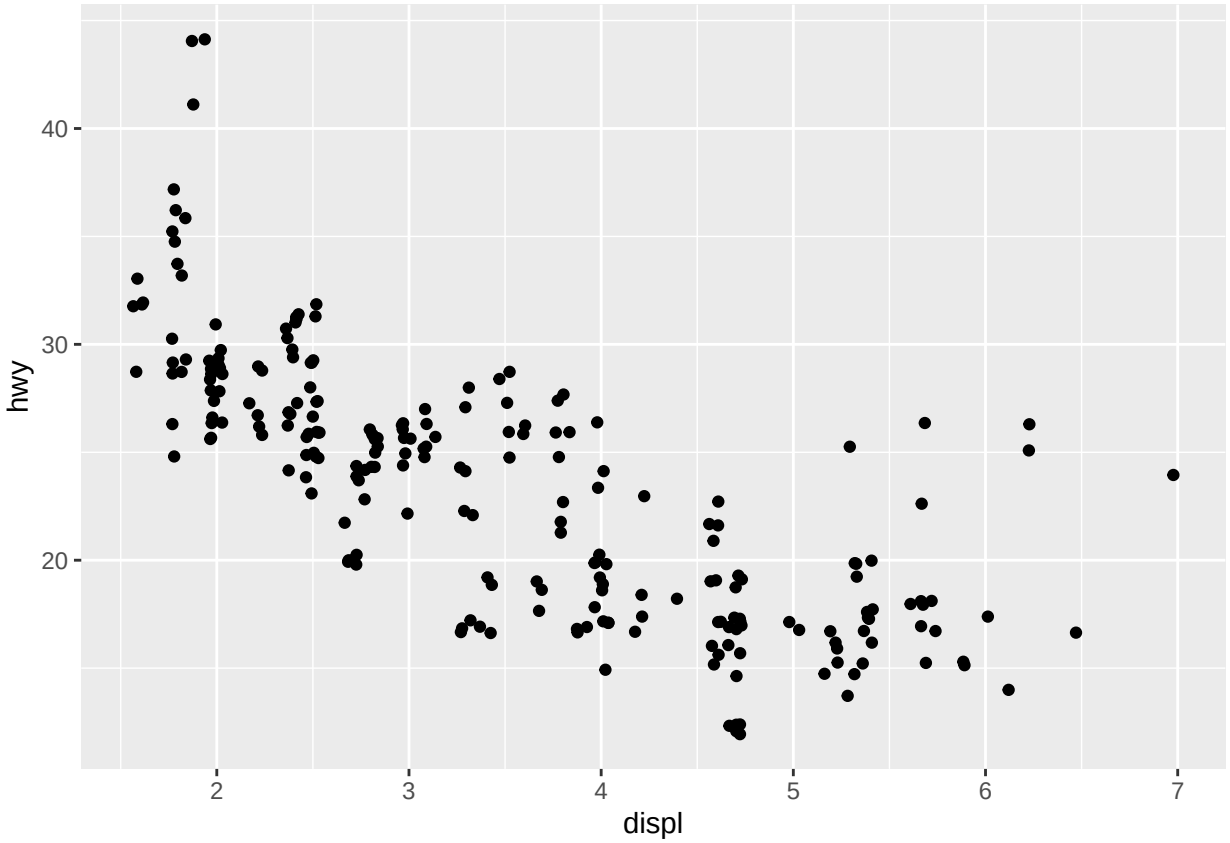

```
ggplot(data = diamonds) +
geom_bar(
mapping = aes(x = cut, fill = clarity),
position = "fill"
)
```



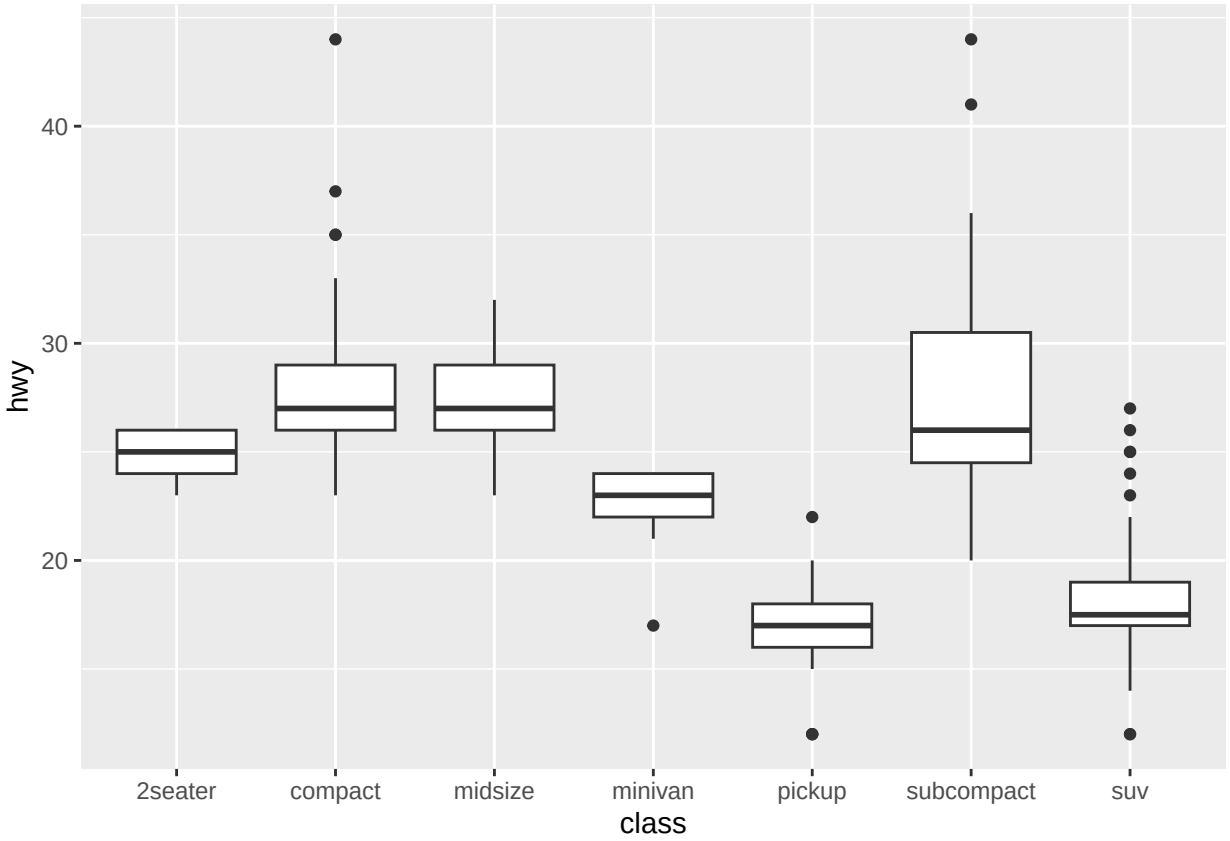

```
ggplot(data = diamonds) +
geom_bar(
mapping = aes(x = cut, fill = clarity),
position = "dodge"
)
```



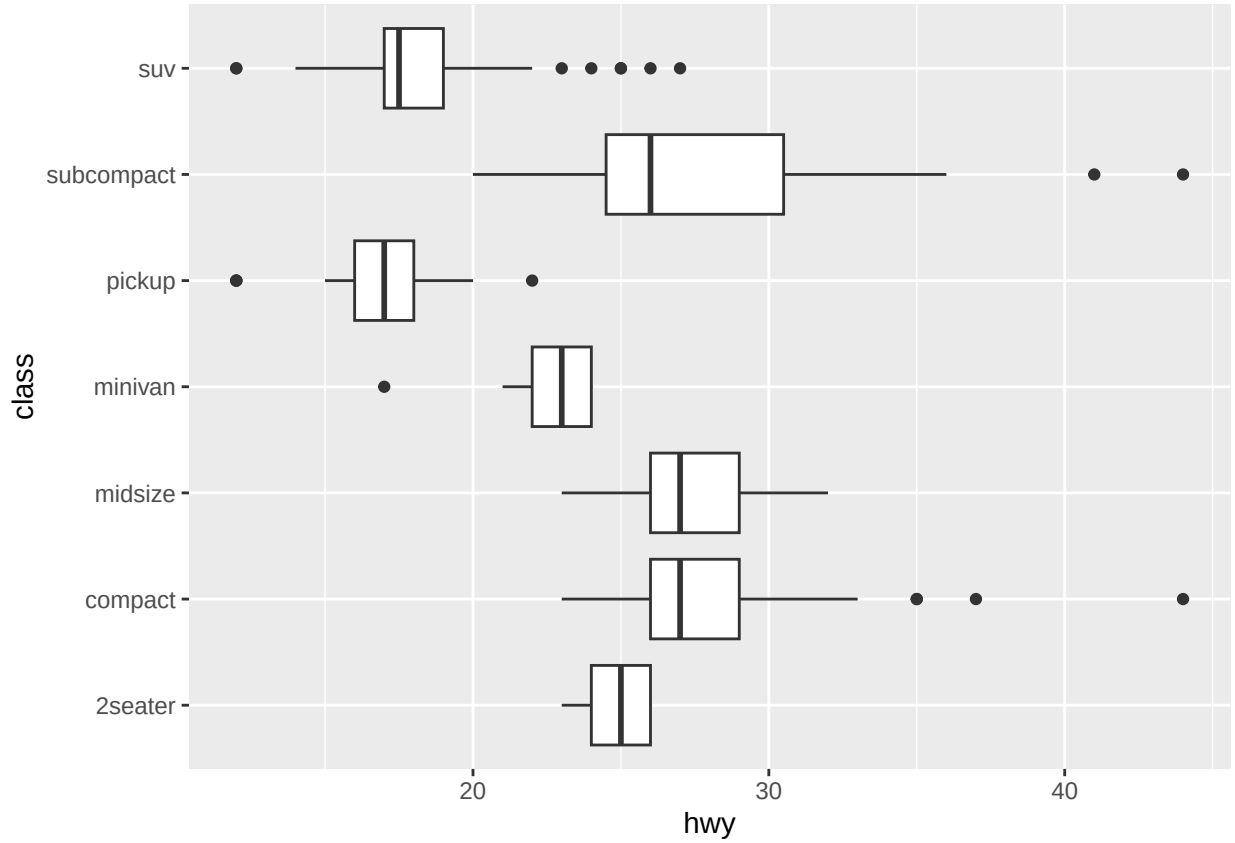

```
ggplot(data = mpg) +
 geom_point(
 mapping = aes(x = displ, y = hwy),
 position = "jitter"
 )
```



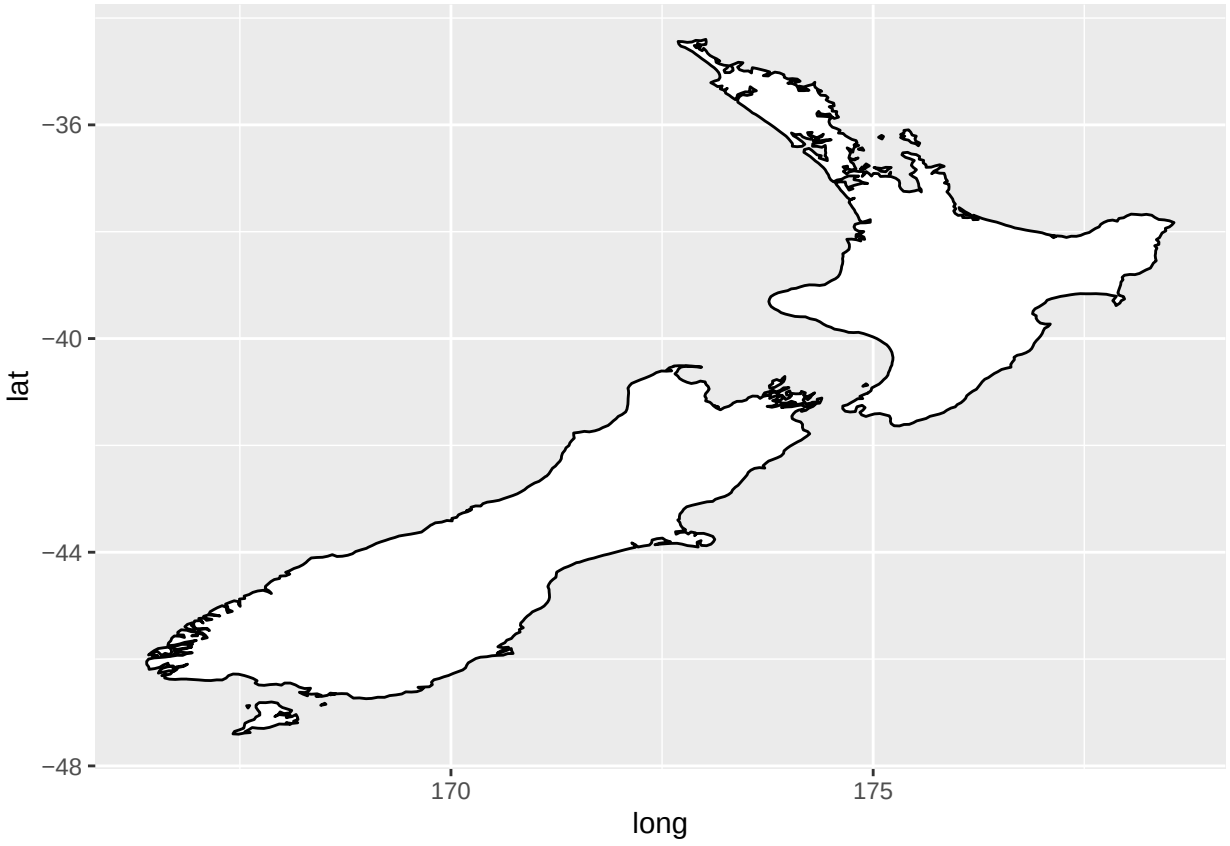

```
ggplot(data = mpg, mapping = aes(x = class, y = hwy)) +
geom_boxplot()
```



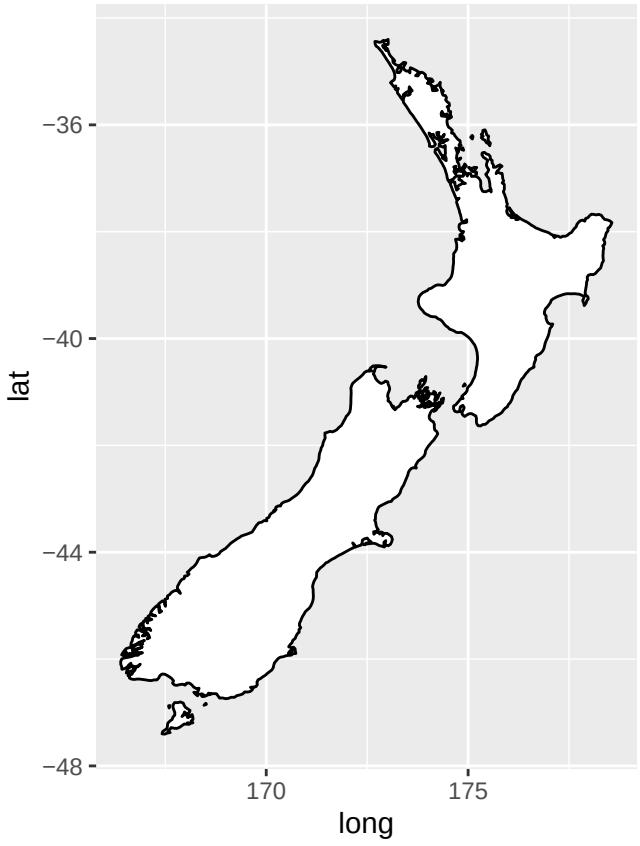

```
ggplot(data = mpg, mapping = aes(x = class, y = hwy)) +
 geom_boxplot() +
 coord_flip()
```



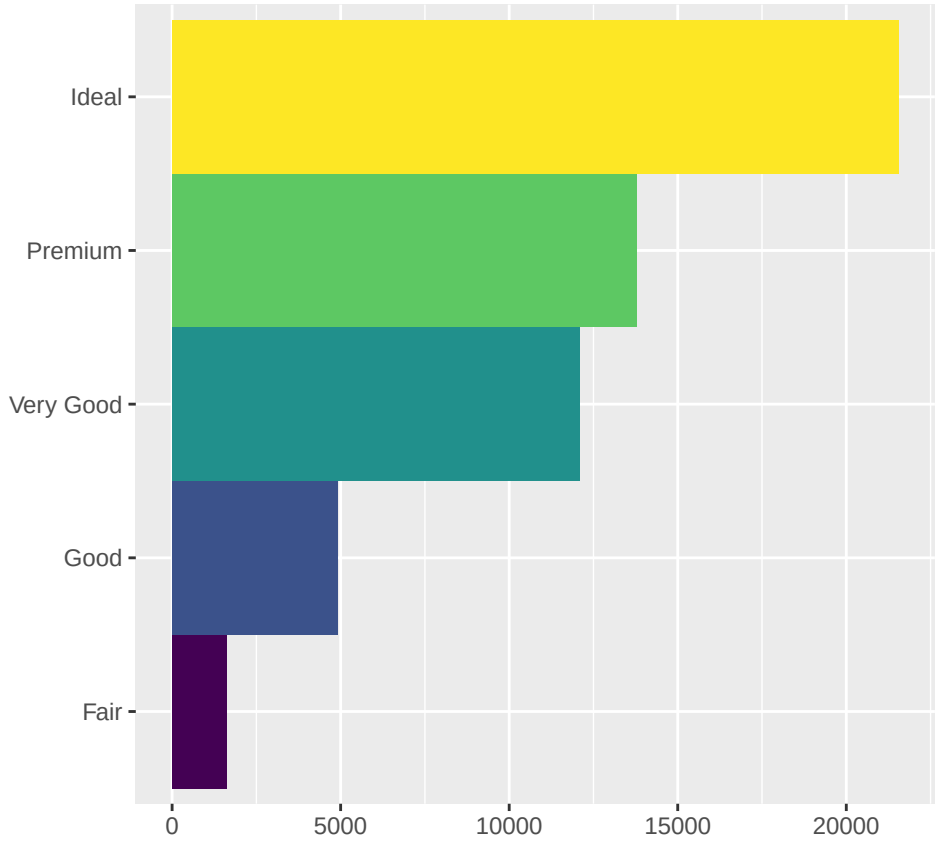

```
nz <- map_data("nz")
ggplot(nz, aes(long, lat, group = group)) +
geom_polygon(fill = "white", color = "black")
```



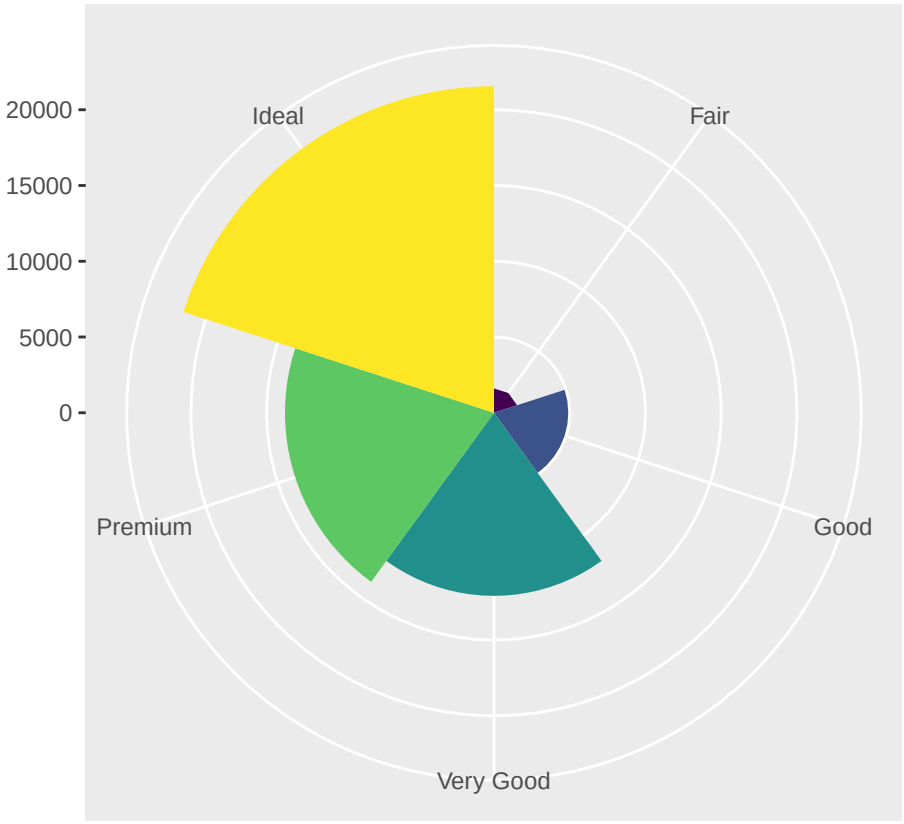

```
ggplot(nz, aes(long, lat, group = group)) +
geom_polygon(fill = "white", color = "black") +
coord_quickmap()
```




```
bar <- ggplot(data = diamonds) +
geom_bar(
mapping = aes(x = cut, fill = cut),
show.legend = FALSE,
width = 1
) +
theme(aspect.ratio = 1) +
labs(x = NULL, y = NULL)
bar + coord_flip()
```




```
bar + coord_polar()
```


20000
15000
10000
5000
0
Ideal
Fair
Premium
Good
Very Good