Data Science for Operational Researchers Using R Online

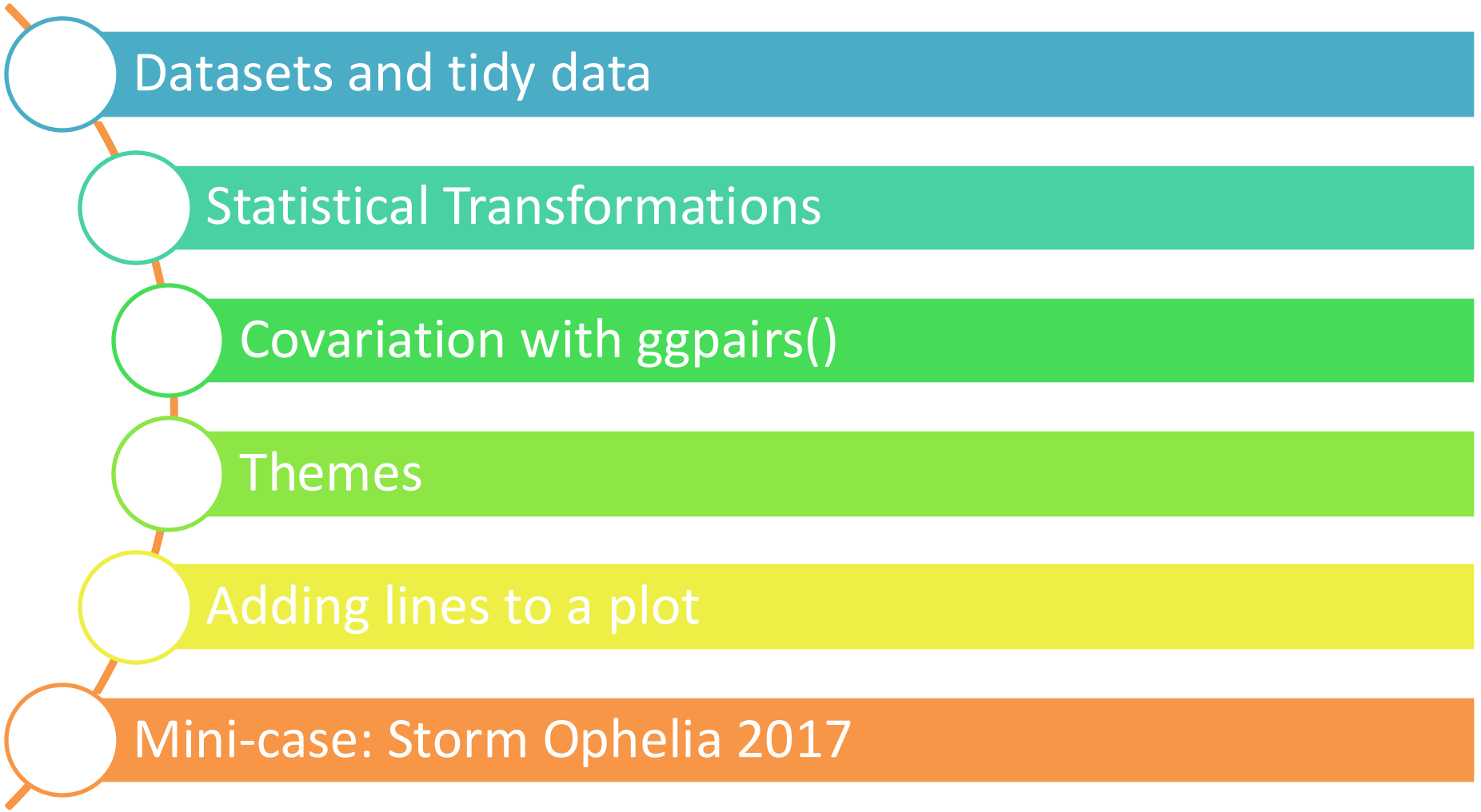
7. Statistical Transformations with ggplot2

Prof. Jim Duggan,
School of Computer Science
University of Galway.

https://github.com/JimDuggan/explore_or

> Data graphics provide one of the most accessible, compelling, and expressive modes to investigate and depict patterns in data.
>
> — Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton (Baumer et al., 2021)

Overview


Datasets and tidy data
Statistical Transformations
Covariation with ggpairs()
Themes
Adding lines to a plot
Mini-case: Storm Ophelia 2017

1. Datasets and tidy data

- For this lecture, we will base our plots on two (tidy) datasets that are already part of the ggplot2 package. These are the tibbles mpg and diamonds.
- With tidy data, where every column is a variable, and every row is an observation.
- These are defined as (Wickham, 2016):
  - A variable is a quantity, quality, or property that you can measure, and will have a value at the time it is measured.
  - An observation is a set of measurements made under similar conditions (often at the same time)

ggplot2::mpg, N = 234

Variable	Description
manufacturer	Manufacturer name
model	Model name
displ	Engine displacement (liters)
year	Year of manufacture
cyl	Number of cylinders
trans	Type of transmission
drv	Type of drive train (e.g. front wheel)
cty	City miles per gallon
hwy	Highway miles per gallon
fl	Fuel type
class	"type" of car (e.g. "compact")

ggplot2::diamonds, N = 53,940

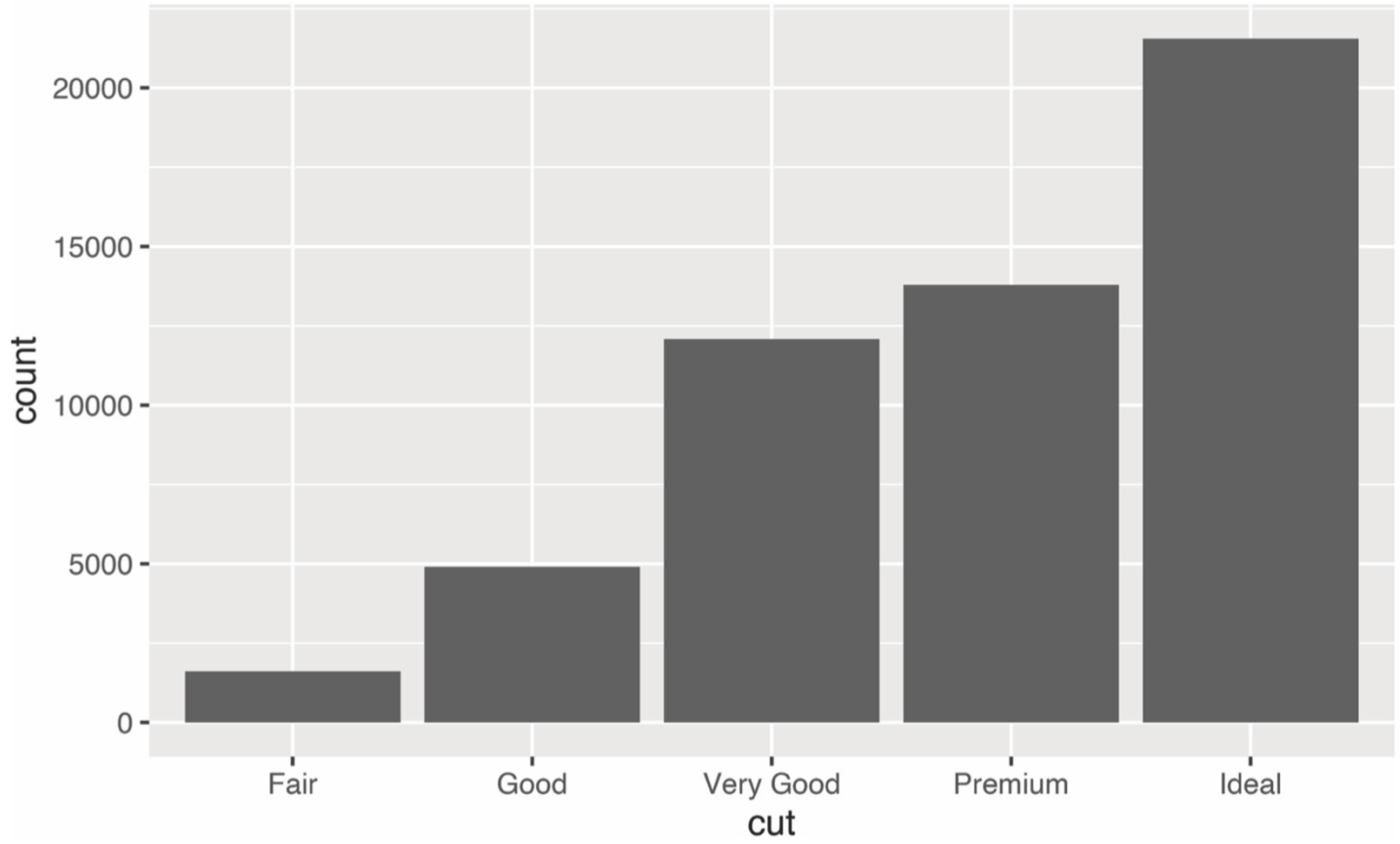
Variable	Description
carat	Weight of the diamond
cut	Quality of the cut (categorical)
color	Diamond color (categorical)
clarity	Diamond clarity (categorical)
depth	Total depth percentage
table	Measure related to width of diamond top
price	Price in dollars
x	Length in mm
y	Width in mm
z	Depth in mm

2. Statistical Transformations

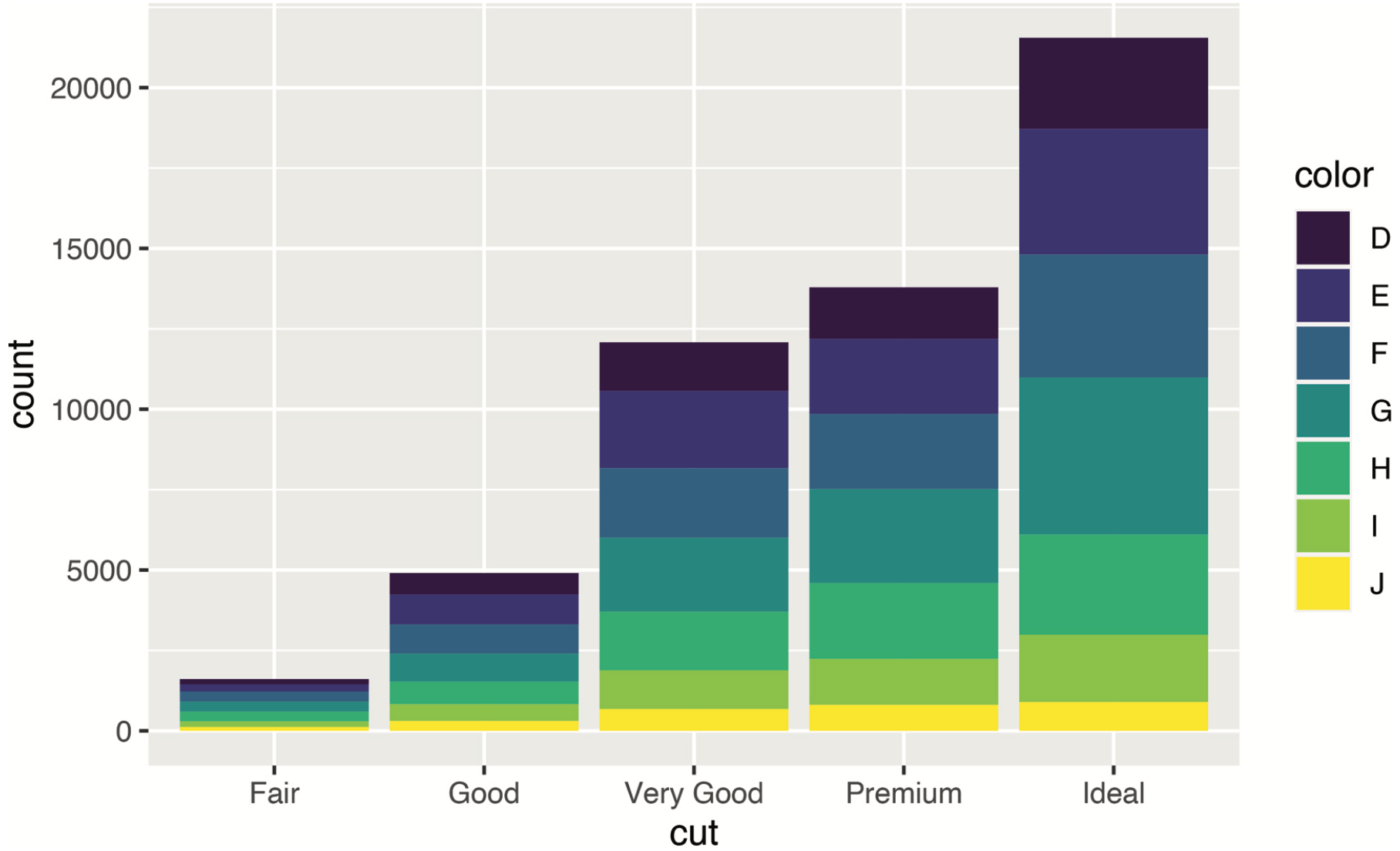
- The scatterplot that we recently explored takes observations from the dataset, and these points are plotted on the graph.
- However, there are graphs that will firstly calculate values, and then plot the results of this calculation
- In ggplot2, this activity is known as what we is termed statistical transformations.

(a) geom_bar() – Frequency count

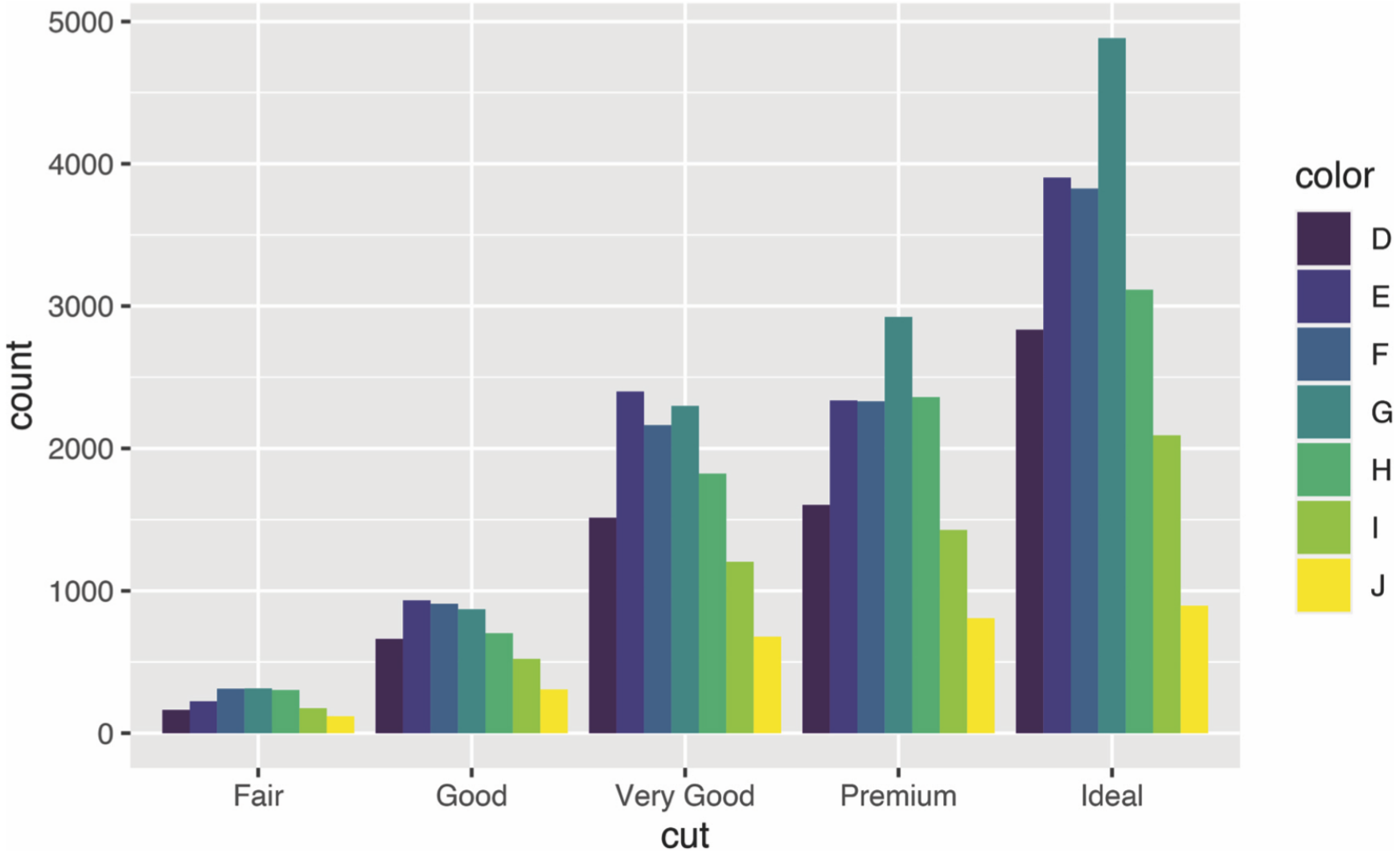
```
ggplot(data=diamonds,mapping=aes(x=cut))+
  geom_bar()
```



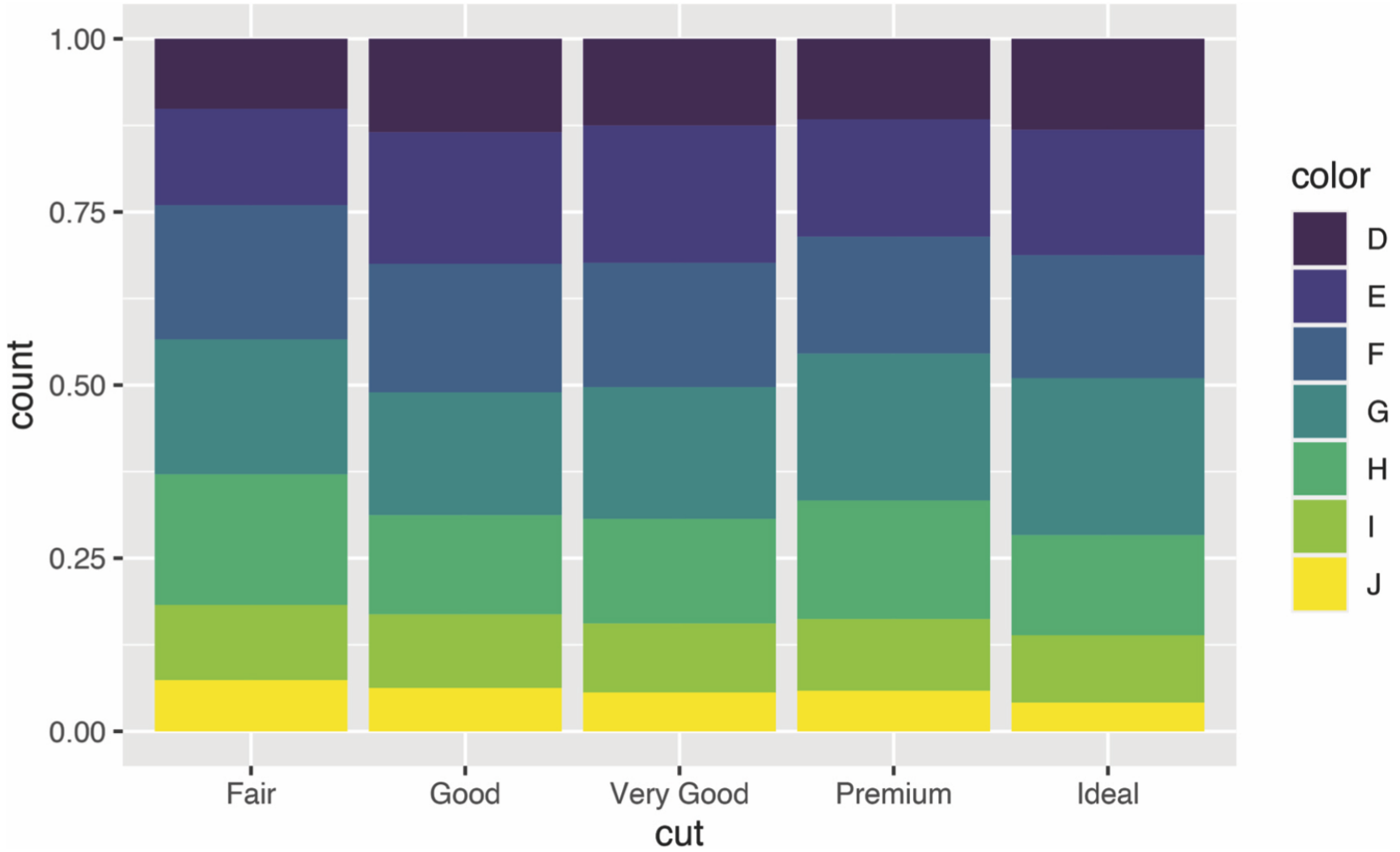
```
ggplot(data=diamonds,mapping=aes(x=cut,fill=color))+
  geom_bar()
```



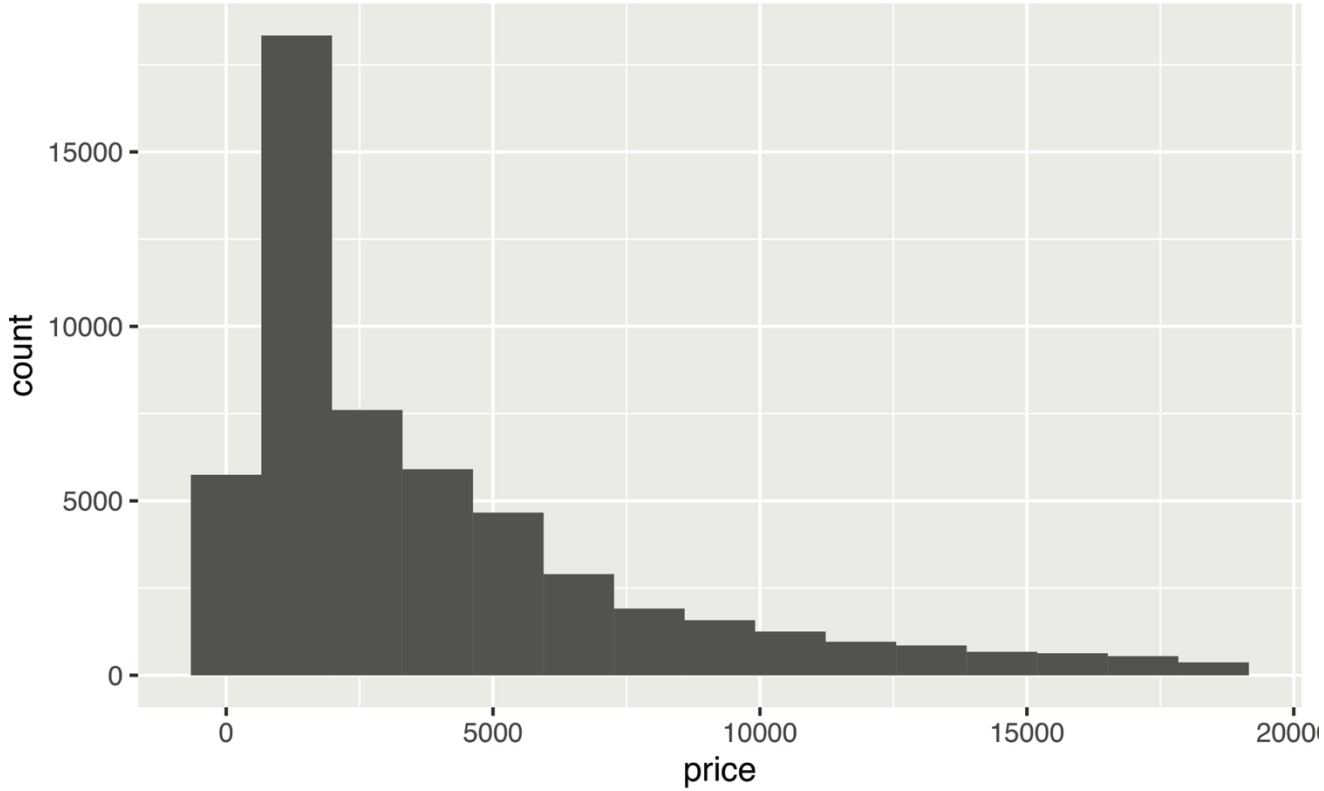
```
ggplot(data=diamonds,mapping=aes(x=cut,fill=color))+
  geom_bar(position="dodge")
```



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



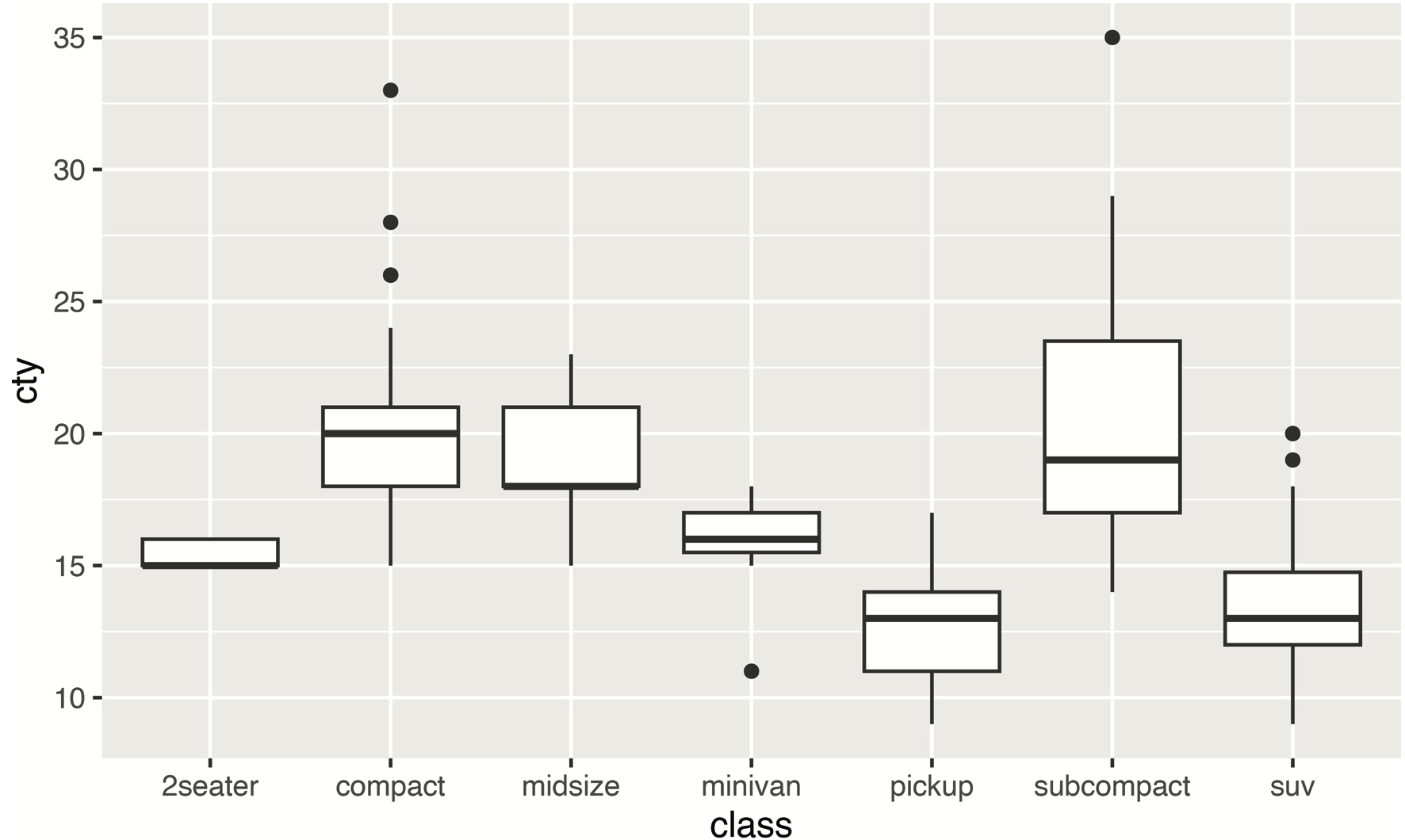
(b) Visualising distributions – geom_histogram()




```
ggplot(data=diamonds,mapping=aes(x=price))+
  geom_histogram(bins = 15)
```

geom_boxplot()

- The boxplot is a valuable graph that visualizes the distribution of a continuous variable, showing the median, the 25th to 75th percentiles, with lines drawn to the position of the value less than or equal to 1.5 times the interquartile range (IQR).
- Values outside 1.5 times the IQR are outliers, and shown as points.

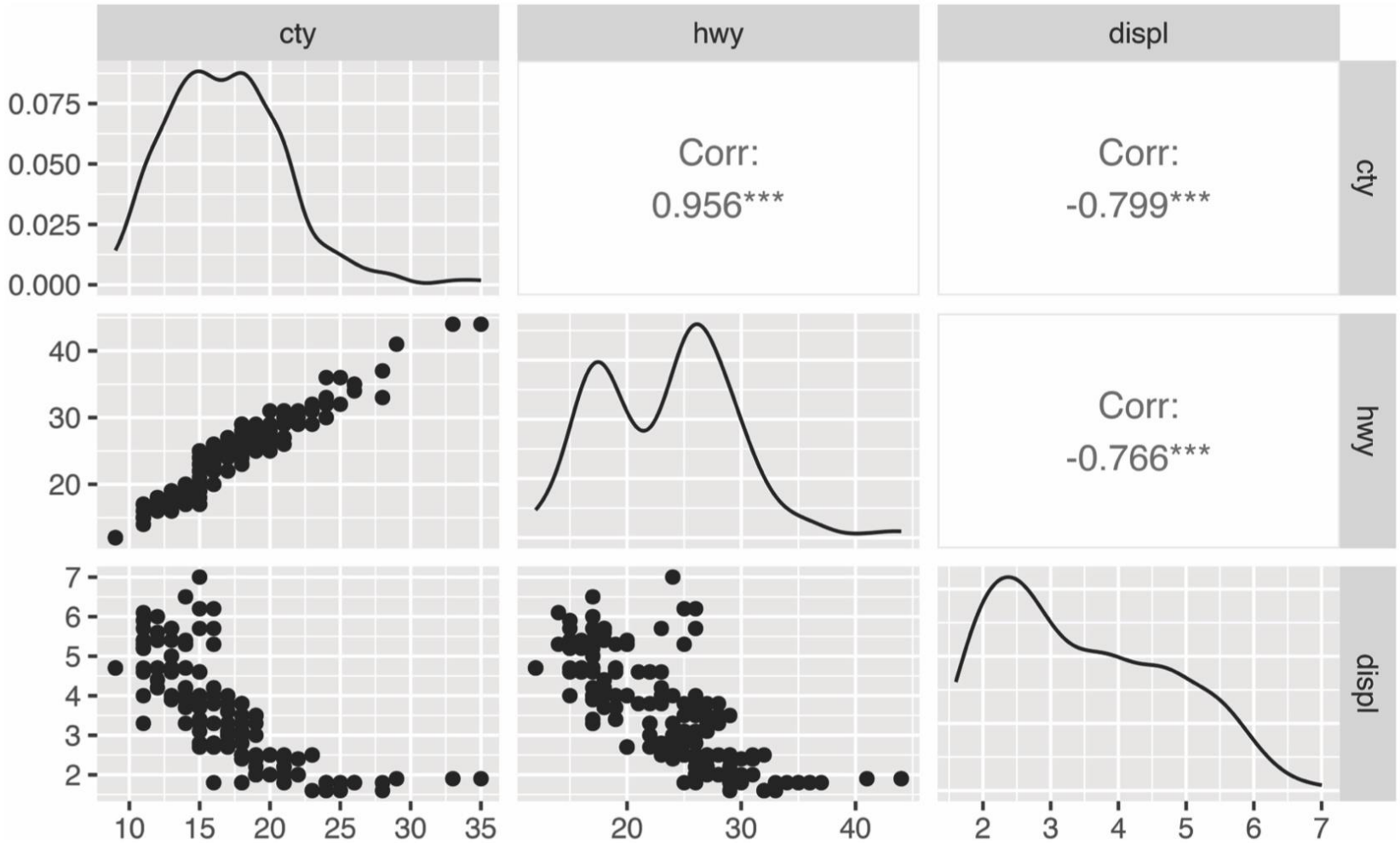



```
ggplot(data=mpg,mapping=aes(y=cty,x=class))+
  geom_boxplot()
```

3. Covariation with ggpairs()

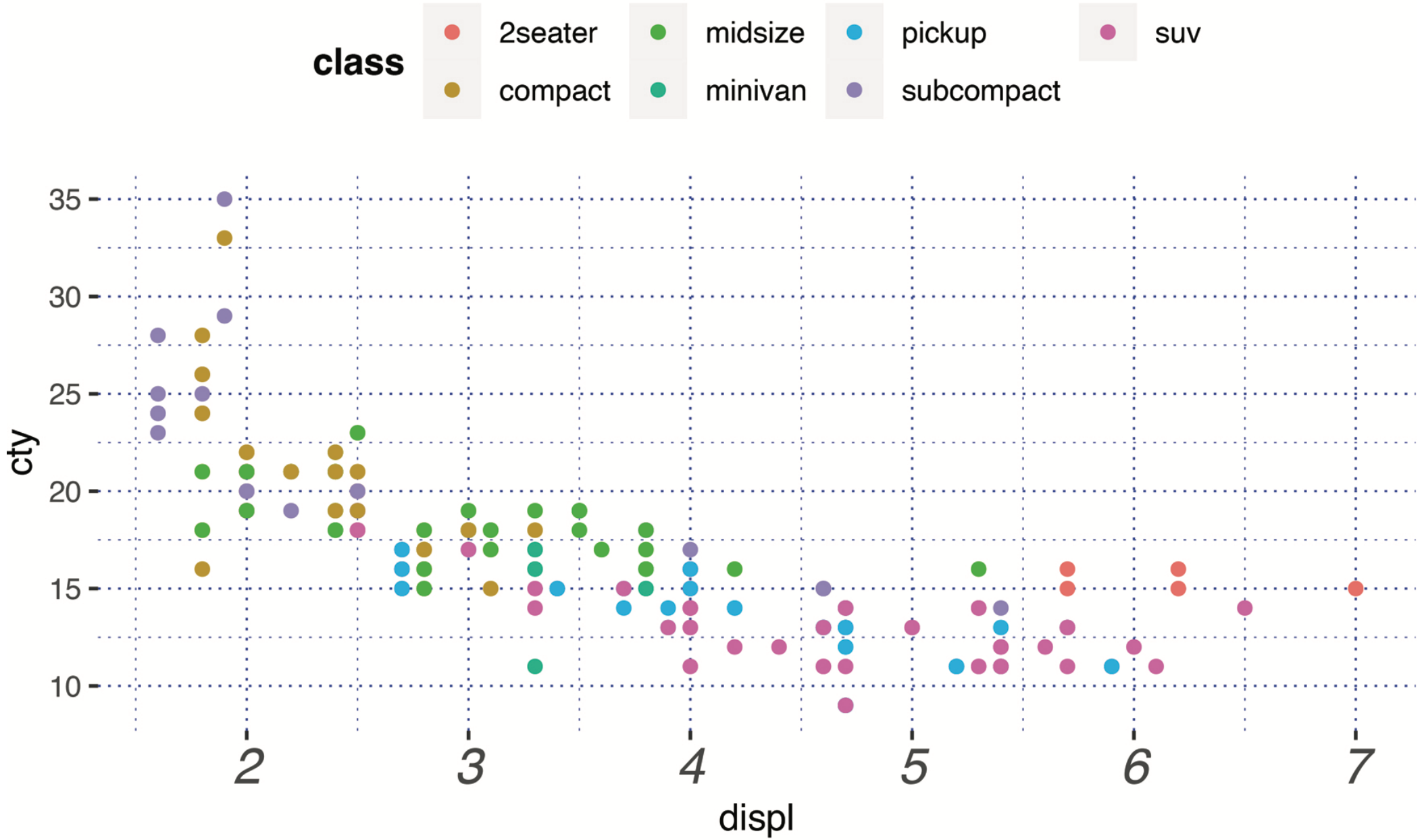
- The package GGally contains the function ggpairs() which visualizes relationships between variables, presents the density plot for each variable, and summarizes the range of correlation coefficients between continuous variables.
- It provides a useful perspective on the data
- For this example, we use the mpg dataset and the three variables cty, hwy, and displ

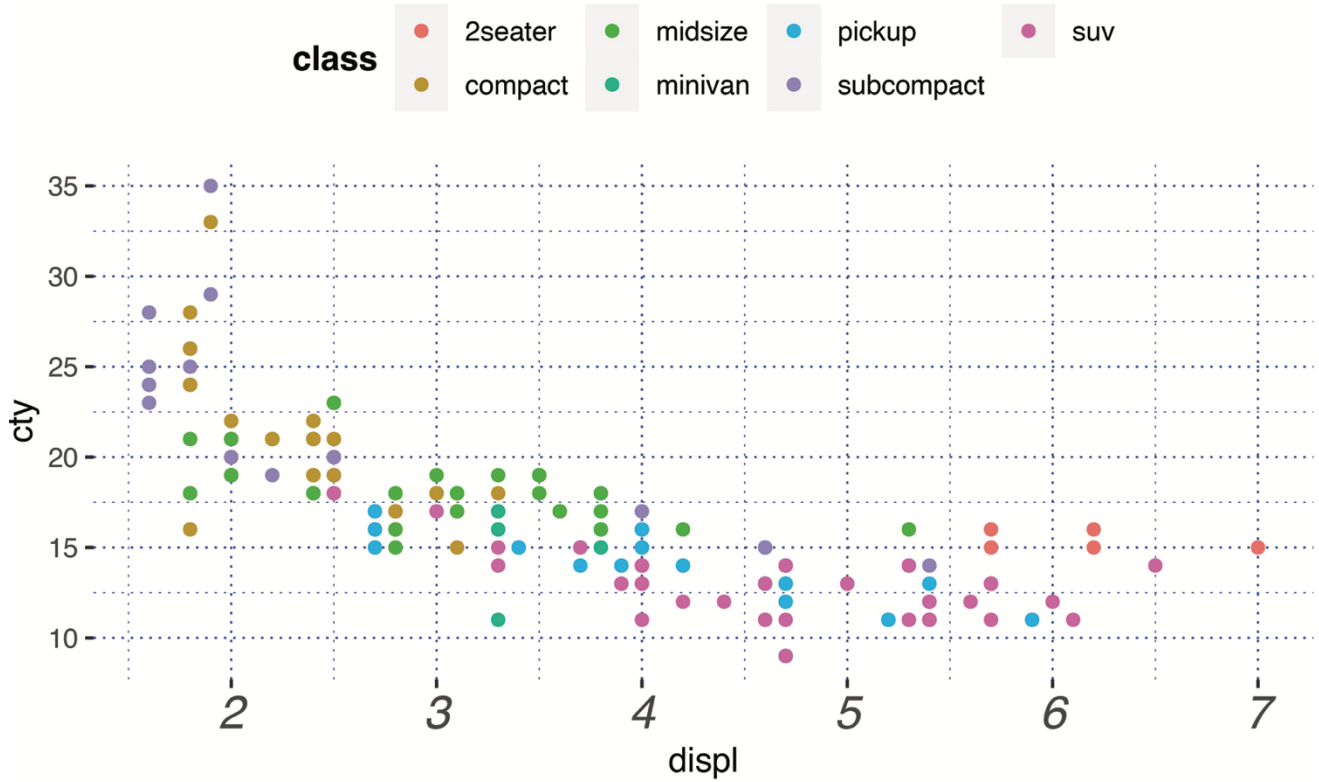
```
library(GGally)
my_vars <- subset(mpg,select=c(cty,hwy,displ))
ggpairs(my_vars)
```



4. Themes

- The ggplot2 theme system enables you to control non-data elements of your plot. For example, control over elements such as fonts, ticks, legends, and backgrounds.
- Components include:
  - Element functions, to describe the visual properties of an element, e.g. element_text() for font size, color, typeface
  - The theme() function to change the default theme elements, e.g. theme(legend.position="top")
  - Complete and ready-to-use themes such as theme_bw()


class
2seater
compact
midsize
minivan
pickup
subcompact
suv
cty
35
30
25
20
15
10
2
3
4
5
6
7
displ



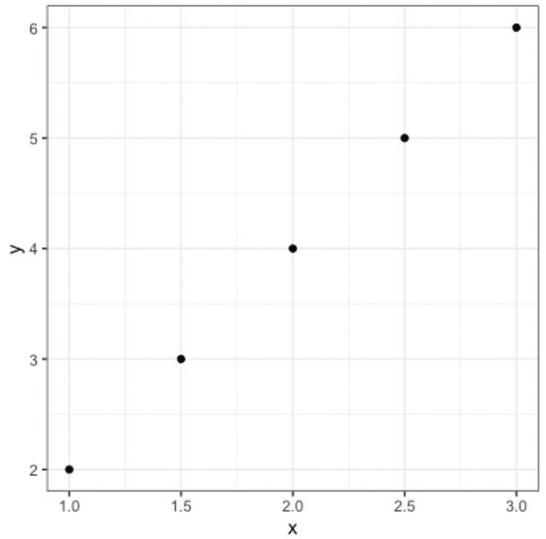

```
ggplot(data=mpg,aes(x=displ,y=cty,color=class))+
  geom_point()+
  theme(legend.title      = element_text(face="bold"),
        legend.position   = "top",
        axis.text.x       = element_text(size=15,face="italic"),
        panel.background  = element_rect(fill="white"),
        panel.grid        = element_line(color = "blue",
                                         linewidth = 0.3,
                                         linetype = 3))
```

Available themes…

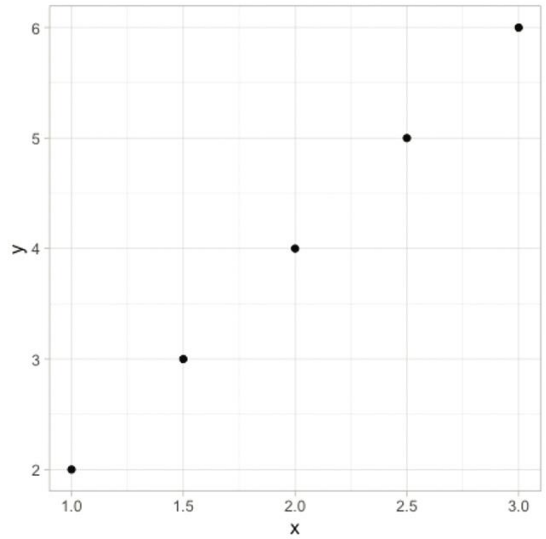
Theme	Description
theme_bw()	a classic dark-on-light theme.
theme_light()	a theme with light gray lines and axes in order to provide more emphasis on the data.
theme_dark()	referred to in the documentation as the dark cousin of theme_light().
theme_minimal()	a minimalistic theme with no background annotations.
theme_classic()	a classic-looking theme, with no gridlines
theme_void()	a completely empty theme.

```
d <- tibble(x=seq(1,3,by=0.5),y=2*x)
plot <- ggplot(data=d,mapping=aes(x=x,y=y))+
          geom_point()
```

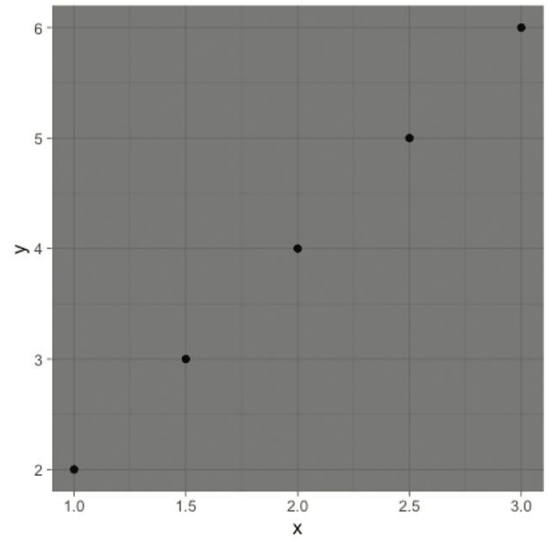
plot+theme_bw()



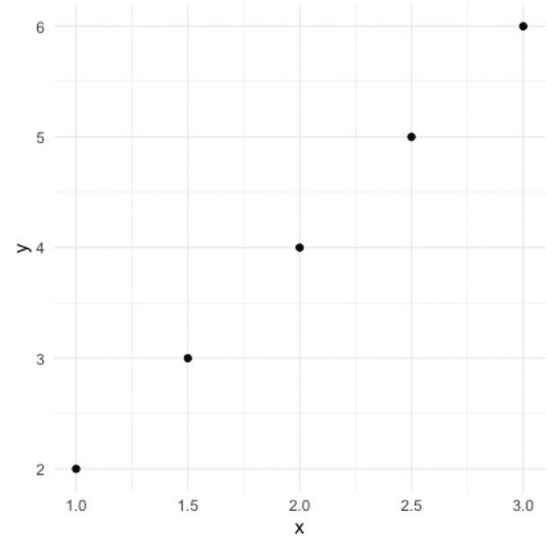
plot+theme_light()



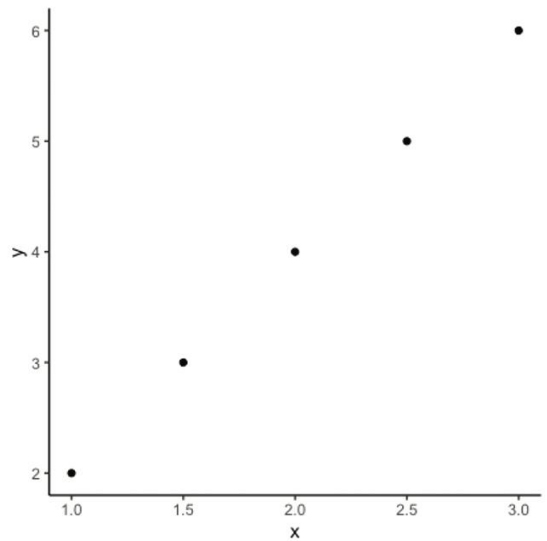
plot+theme_dark()



plot+theme_minimal()



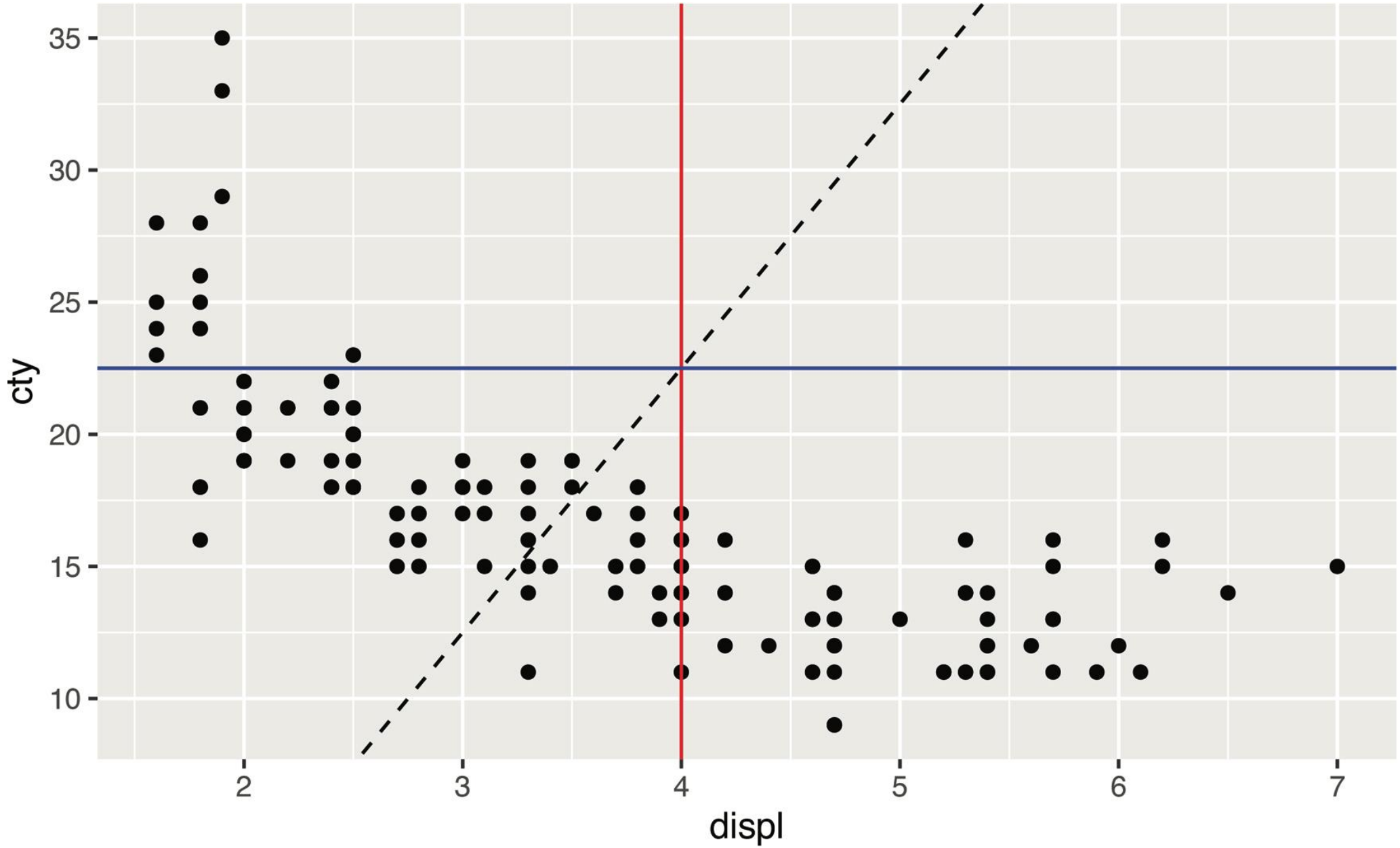
plot+theme_classic()

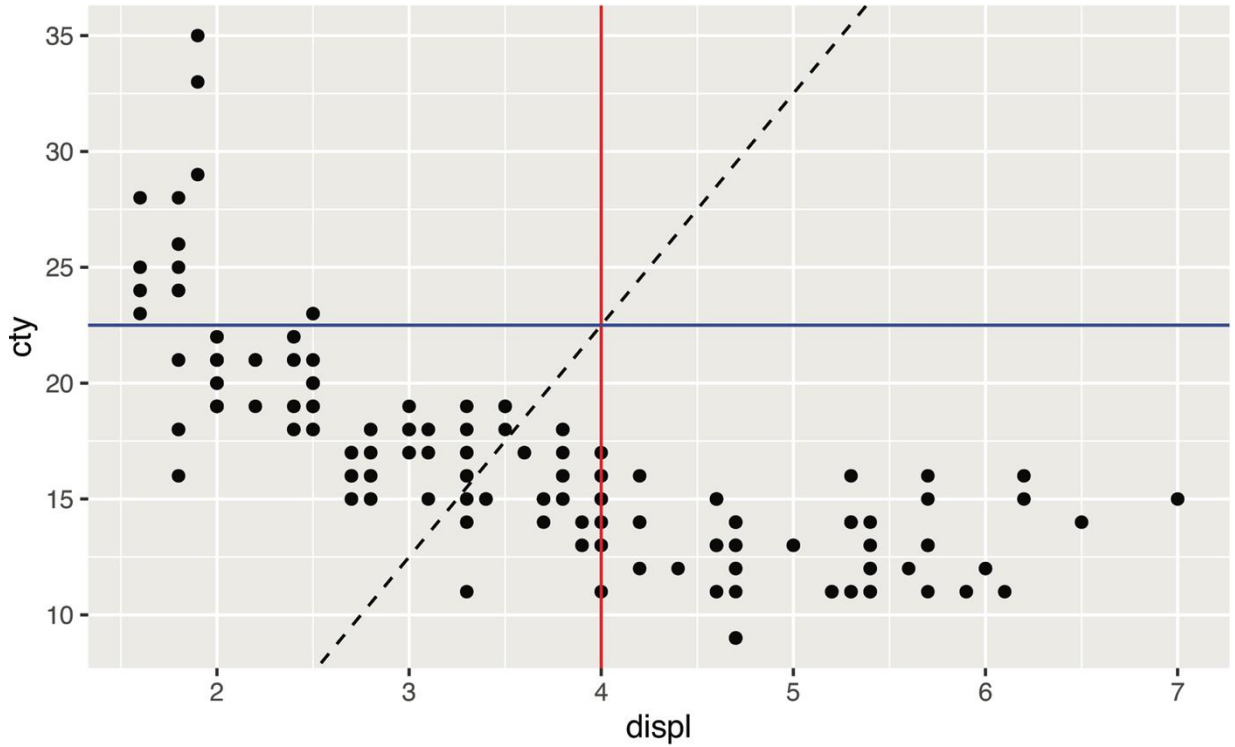


plot+theme_void()

5. Adding lines to a plot

- There may be situations where adding a layer containing a line is required for your plot.
- There are three functions for doing this, specifically:
  - geom_vline(), where the argument xintercept is fixed to position the vertical line.
  - geom_hline(), with the argument yintercept used to position the horiztonal line.
  - geom_abline(), where both the arguments slope and intercept can be set to generate the output line.


35
30
25
20
15
10
cty
2
3
4
5
6
7
displ




```
ggplot(data=mpg,mapping=aes(x=displ,y=cty))+
  geom_point()+
  geom_vline(xintercept = 4,color="red")+
  geom_hline(yintercept = 22.5,color="blue")+
  geom_abline(slope=10,intercept=-17.5,linetype="dashed")
```

6. Mini-case – Storm Ophelia 2017

- We take our first look at the dataset aimsir17, which contains hourly observations from 25 weather stations in Ireland, all recorded in 2017
- There are 219,000 records in all (25 stations × 365 days × 24 hourly observations)
- Data recorded includes the month, day, hour, time, rainfall, temperature, humidity, mean sea level pressure, wind speed, and wind direction.
- Note that the date variable is a special R data format that is recognized by ggplot2 for time series plots

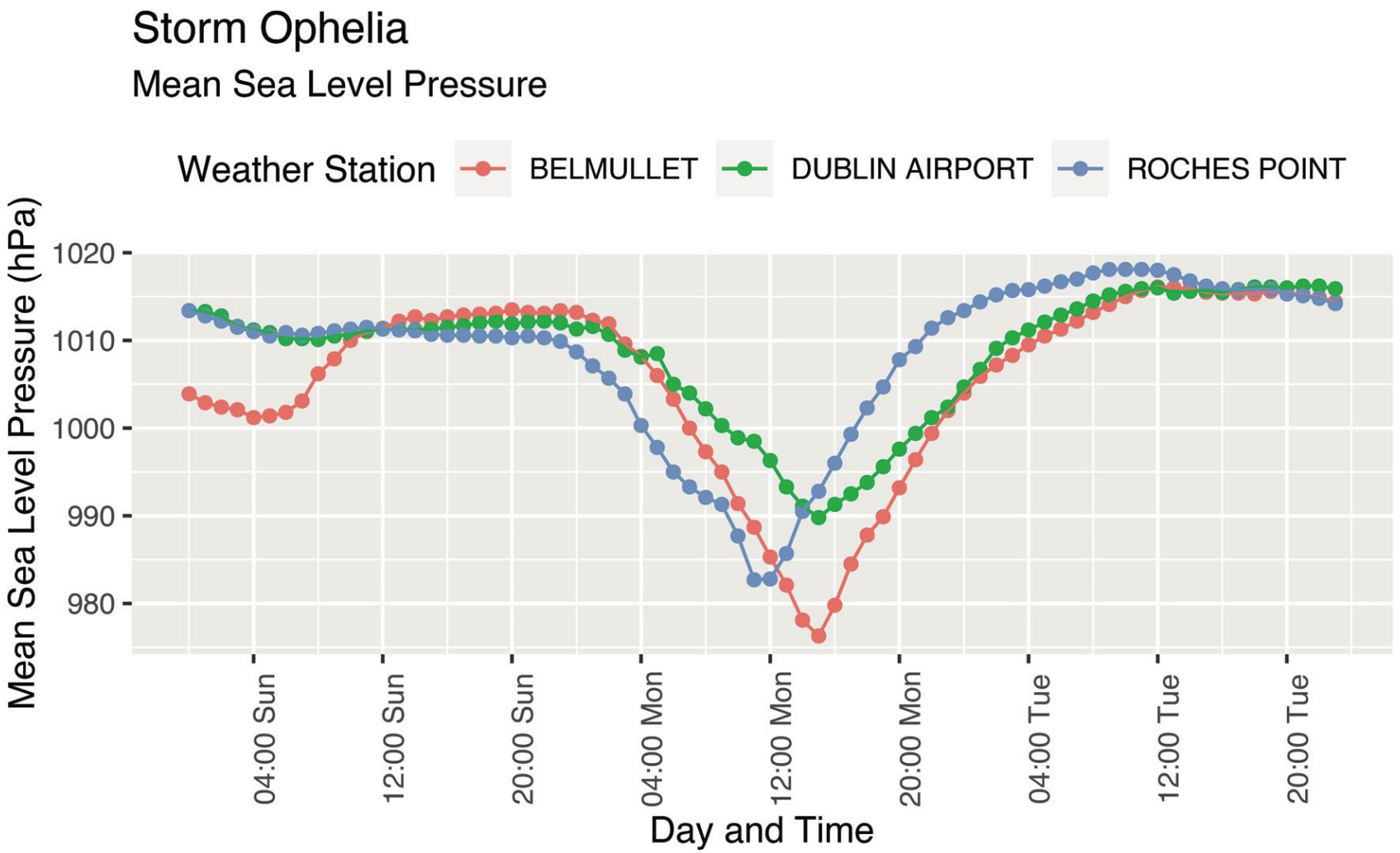
```
library(aimsir17)

observations
#> # A tibble: 219,000 x 12
#>    station  year month   day  hour date                 rain
#>    <chr>   <dbl> <dbl> <int> <int> <dttm>              <dbl>
#>  1 ATHENRY  2017     1     1     0 2017-01-01 00:00:00   0
#>  2 ATHENRY  2017     1     1     1 2017-01-01 01:00:00   0
#>  3 ATHENRY  2017     1     1     2 2017-01-01 02:00:00   0
#>  4 ATHENRY  2017     1     1     3 2017-01-01 03:00:00   0.1
#>  5 ATHENRY  2017     1     1     4 2017-01-01 04:00:00   0.1
#>  6 ATHENRY  2017     1     1     5 2017-01-01 05:00:00   0
#>  7 ATHENRY  2017     1     1     6 2017-01-01 06:00:00   0
#>  8 ATHENRY  2017     1     1     7 2017-01-01 07:00:00   0
#>  9 ATHENRY  2017     1     1     8 2017-01-01 08:00:00   0
#> 10 ATHENRY  2017     1     1     9 2017-01-01 09:00:00   0
#> # ... with 218,990 more rows, and 5 more variables: temp <dbl>,
#> #   rhum <dbl>, msl <dbl>, wdsp <dbl>, wddir <dbl>
```

Data preparation

- We first need to “drill down” and generate a reduced dataset for the three days in question, and we will also select a number of weather stations from different parts of Ireland.
- These are Belmullet (north west), Roches Point (south), and Dublin Airport (east).
- The following code uses the function subset() to filter the dataset, and the & operator is used as we have more than one filtering condition.
- We also use the %in% operator, which is convenient

```
storm <- observations |>
           subset(month==10 &
                  day %in% 15:17 &
                  station %in% c("BELMULLET",
                                 "ROCHES POINT",
                                 "DUBLIN AIRPORT"),
                  select=c(station,date,temp,msl,wdsp))
storm
#> # A tibble: 216 x 5
#>    station   date                 temp   msl  wdsp
#>    <chr>     <dttm>              <dbl> <dbl> <dbl>
#>  1 BELMULLET 2017-10-15 00:00:00  15.5 1004.    30
#>  2 BELMULLET 2017-10-15 01:00:00  15.9 1003.    33
#>  3 BELMULLET 2017-10-15 02:00:00  15.3 1002.    33
#>  4 BELMULLET 2017-10-15 03:00:00  15.1 1002.    35
#>  5 BELMULLET 2017-10-15 04:00:00  15.1 1001.    33
```

Exploration 1: Mean sea level pressure



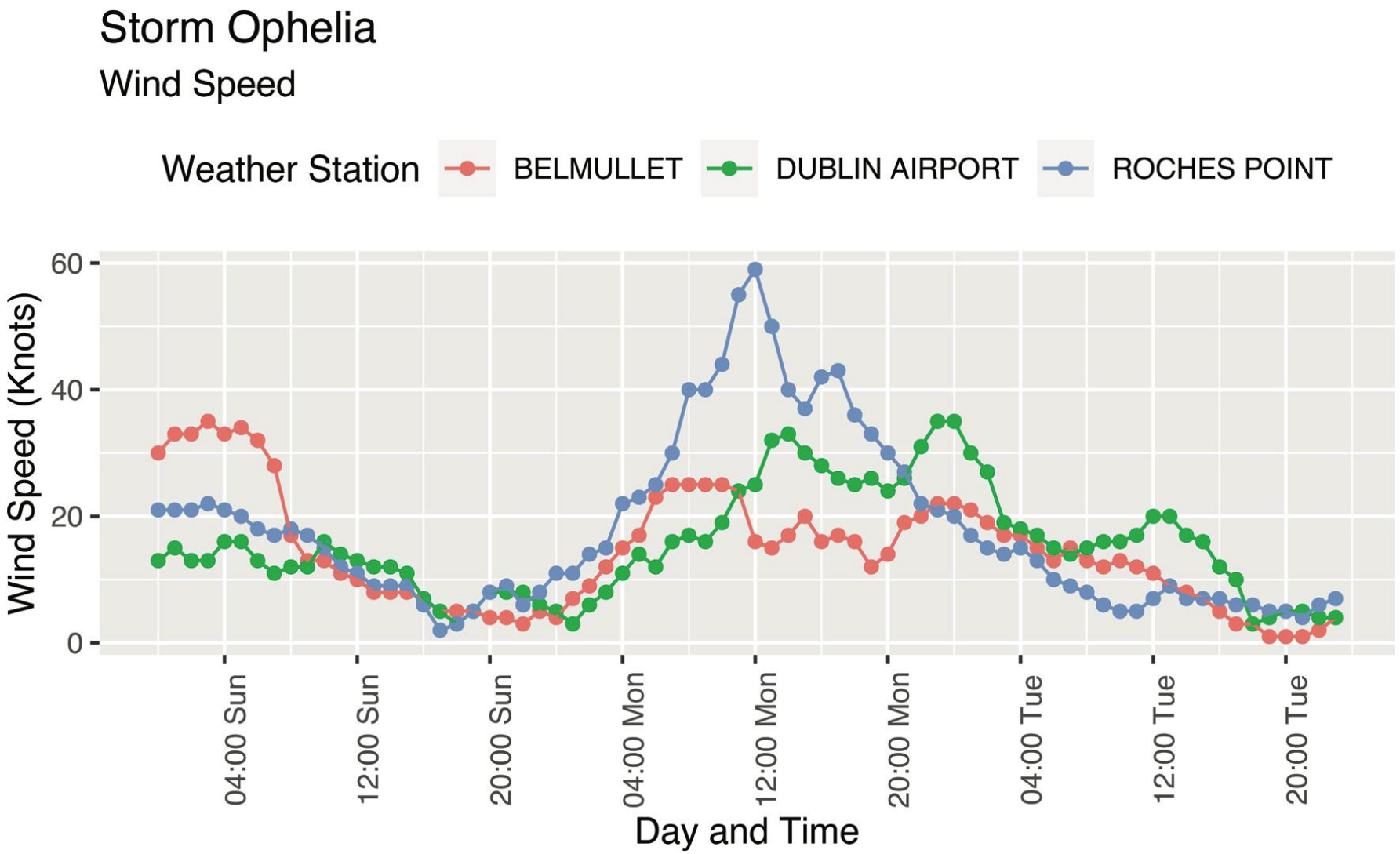
```
ggplot(data=storm,aes(x=date,y=msl,color=station))+
  geom_point()+
  geom_line() +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90))+
  scale_x_datetime(date_breaks = "8 hour",date_labels = "%H:%M %a")+
  labs(title="Storm Ophelia",
       subtitle = "Mean Sea Level Pressure",
       x="Day and Time",
       y="Mean Sea Level Pressure (hPa)",
       color="Weather Station")
```

scale_x_datetime()

- scale_x_datetime() provides an excellent way to format dates, and has two arguments
  - date_breaks which specifies the x-axis points that will contain date labels (every 8 hours)
  - date_labels which allows you to configure what will be printed (hour, minute, and day of week).
  - '%H:%M %a" will combine the hour (24-hour clock), a colon, the minute (00–59), a blank space followed by the abbreviated day of the week.

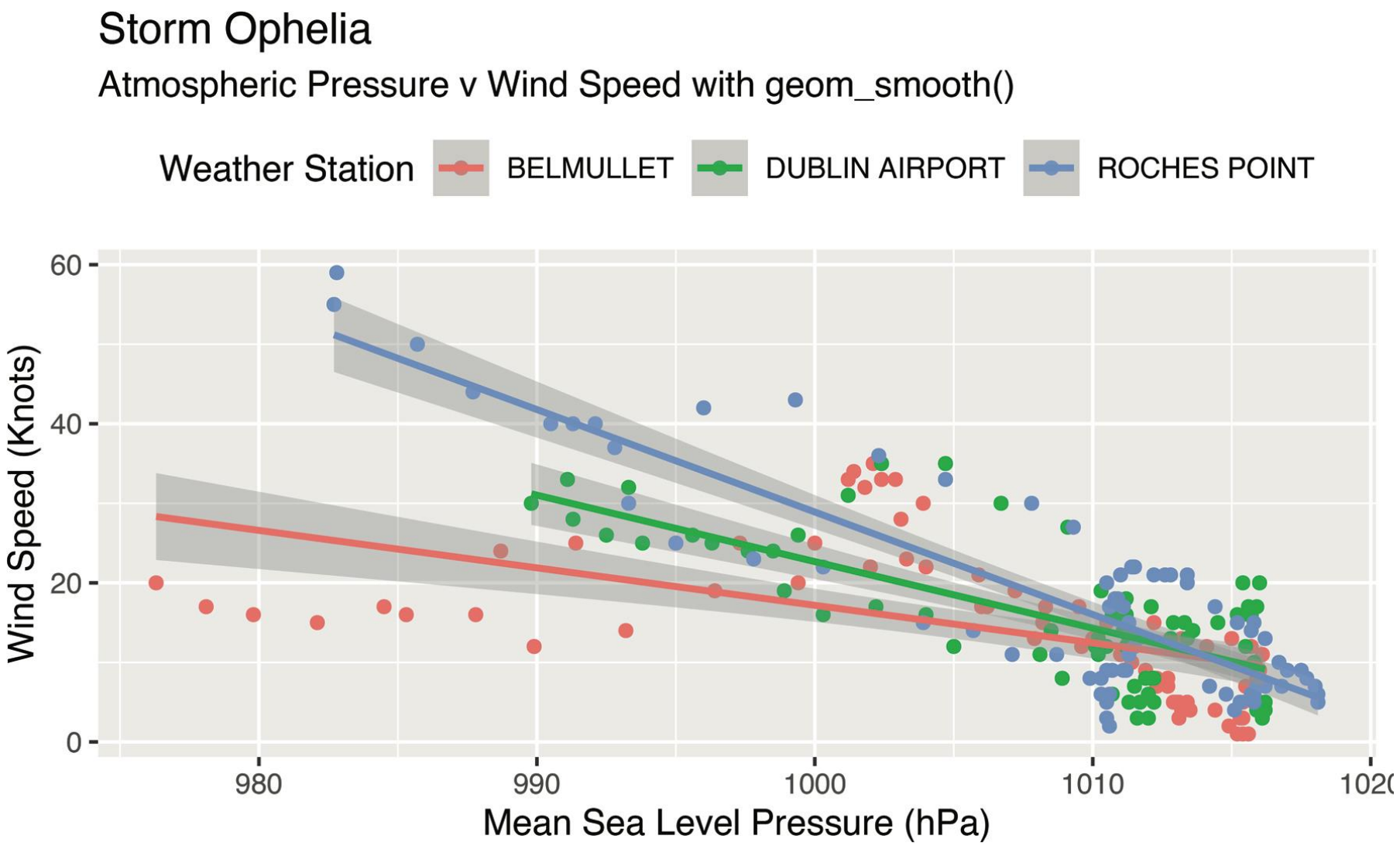
String	Meaning for `date_labels`
%S	Second (00–59)
%M	Minute (00–59)
%l	Hour, 12 hour clock (1–12)
%I	Hour, 12 hour clock (01–12)
%p	am/pm
%H	Hour, 24-hour clock (00–23)
%a	Day of week, abbreviated (Mon–Sun)
%A	Day of week, full (Monday–Sunday)
%e	Day of month (1–31)
%d	Day of month (00–31)
%m	Month, numeric (01–12)
%b	Month, abbreviated (Jan–Dec)
%B	Month, full (January–December)
%y	Year, without century (00–99)
%Y	Year, with century (0000–9999)

Exploration 2: Average hourly wind speed



```
ggplot(data=storm,aes(x=date,y=wdsp,color=station))+
  geom_point()+
  geom_line()+
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90))+
  scale_x_datetime(date_breaks = "8 hour",date_labels = "%H:%M %a")+
  labs(title="Storm Ophelia",
       subtitle = "Wind Speed",
       x="Day and Time",
       y="Wind Speed (Knots)",
       color="Weather Station")
```

Exploration 3: Associations between MSL and wind



```
ggplot(data=storm,aes(x=msl,y=wdsp,color=station))+
  geom_point()+
  geom_smooth(method="lm")+
  theme(legend.position = "top")+
  labs(title="Storm Ophelia",
       subtitle="Atmospheric Pressure v Wind Speed with geom_smooth()",
       x="Mean Sea Level Pressure (hPa)",
       y="Wind Speed (Knots)",
       color="Weather Station")
```

2. Use `ggpairs()` to explore a subset of the Irish weather dataset (`aimsir17`), and the possible relationships between the variables `rain`, `wdsp`, `temp` and `msl`. Use the station "ROCHES POINT" and all the observations from the month of October (month 10), which can be retrieved using the `subset()` function.