intro to tidyverse

ryan snoyman

install

a) Download R
https://cran.rstudio.com/

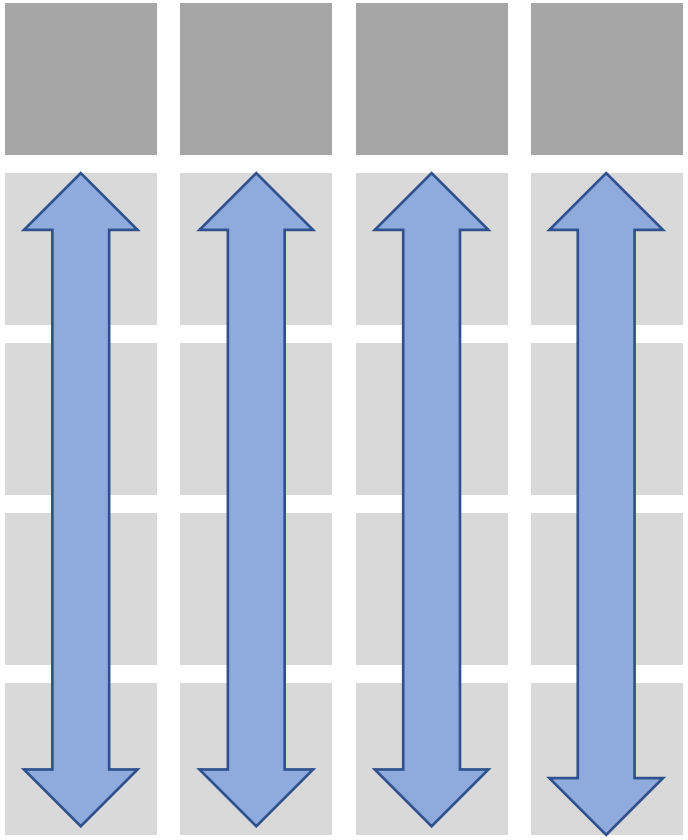
b) Download R studio IDE
https://rstudio.com/products/rstudio/download/

c) Install tidyverse

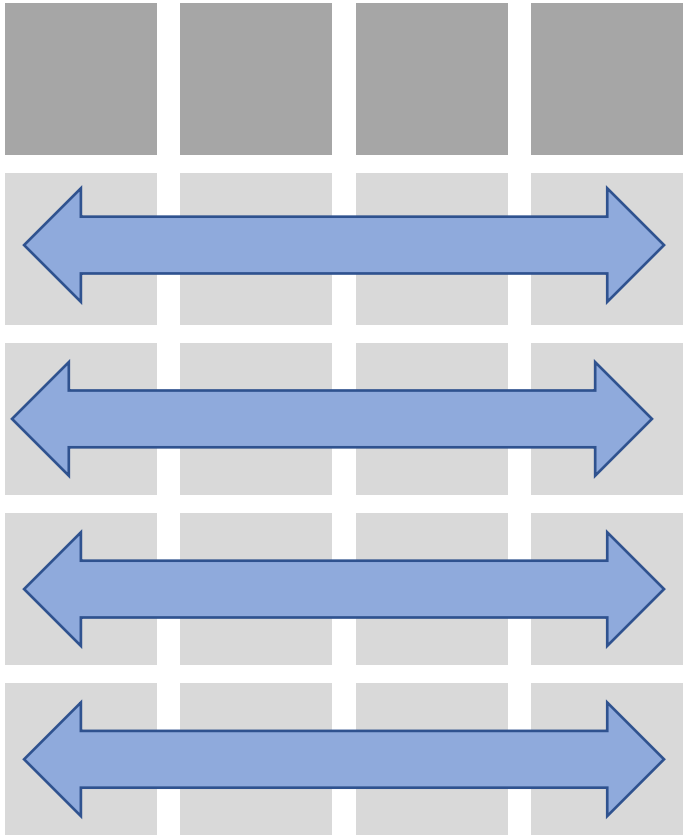
```
install_packages('tidyverse')
library(tidyverse)
```

tidy data

variables



observations

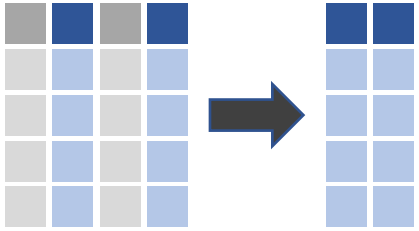


diamonds

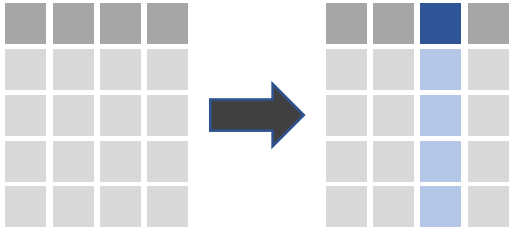
carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326		3.98	2.43
0.23	Good	E	VS1	56.9	65	327		4.07	2.31
0.29	Premium	I	VS2	62.4	58	334		4.23	2.63
0.31	Good	J	SI2	63.3	58	335		4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336		3.96	2.48

dplyr

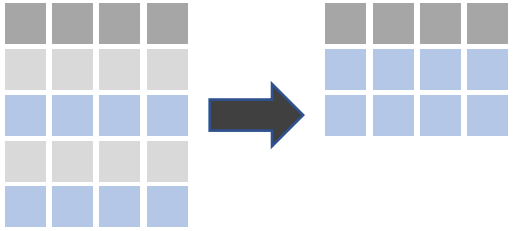
select()



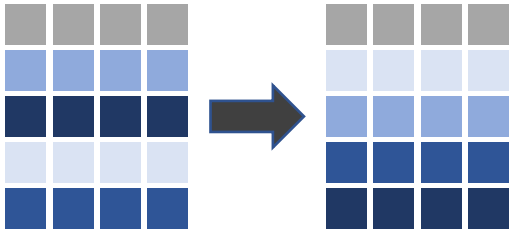
mutate()



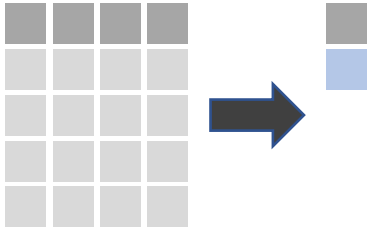
filter()



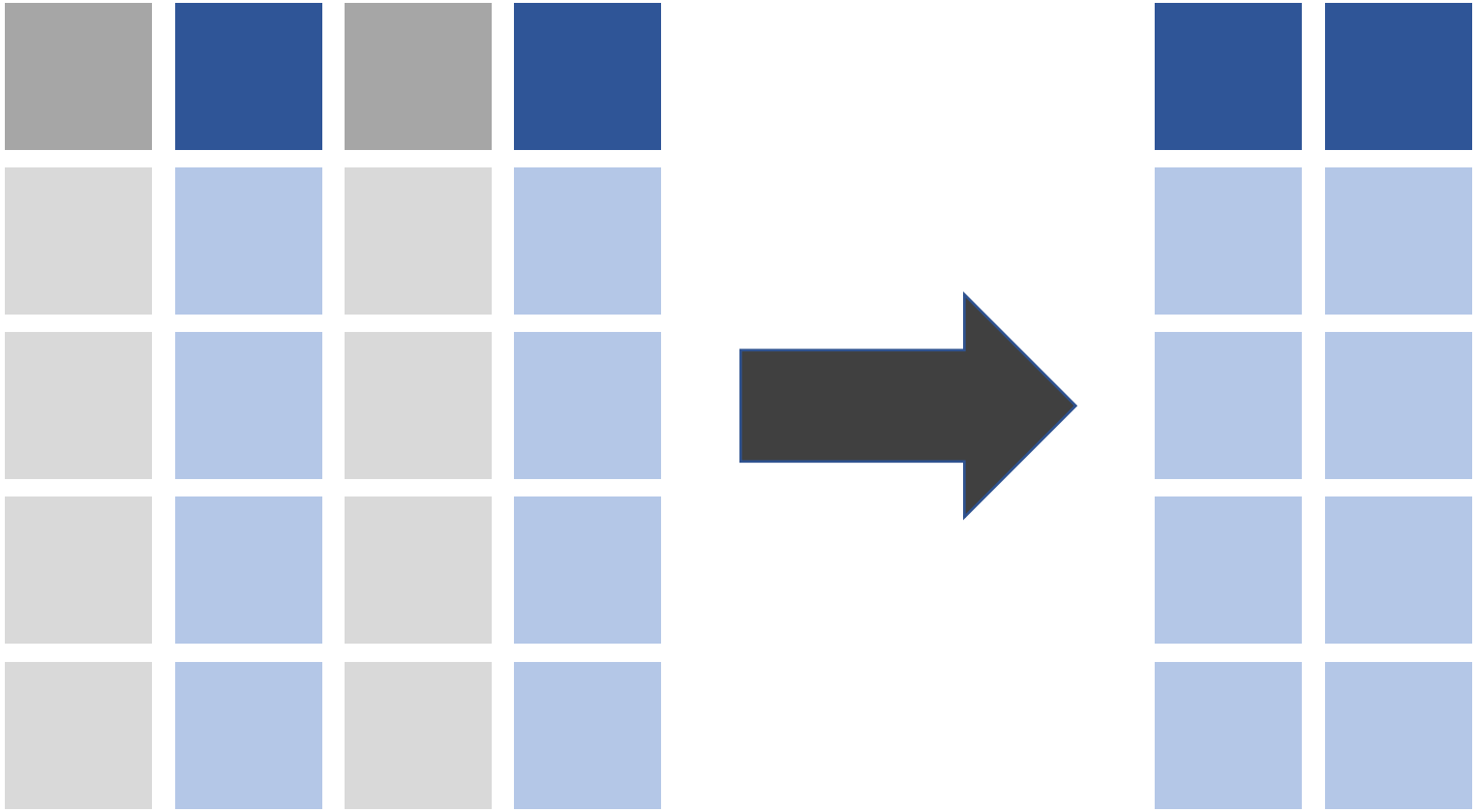
arrange()



summarise()

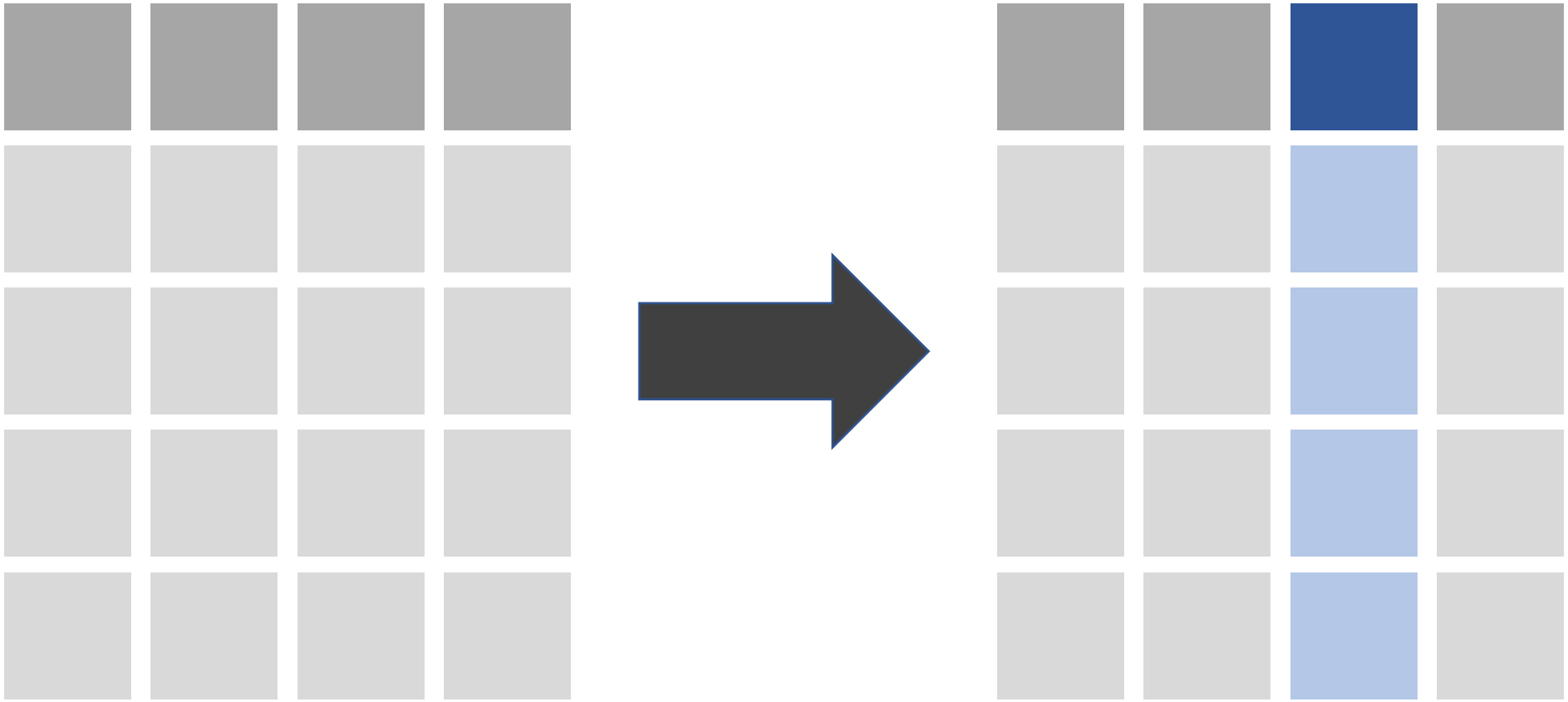


select()



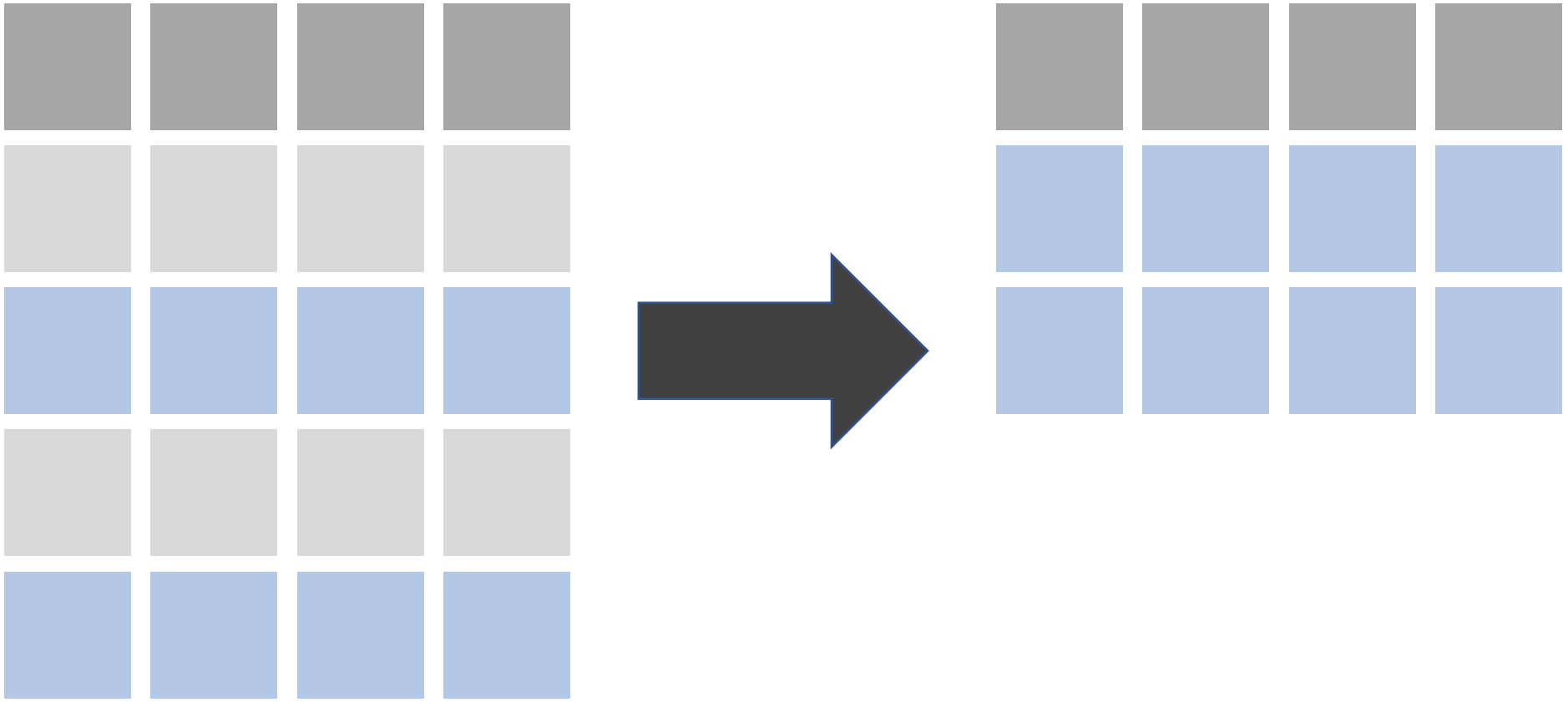
```
select(diamonds, carat, cut, color, clarity, price)

select(diamonds, carat:clarity, price)

select(diamonds, 1:4, price)

select(diamonds, starts_with(‘c’), price)
```

mutate()



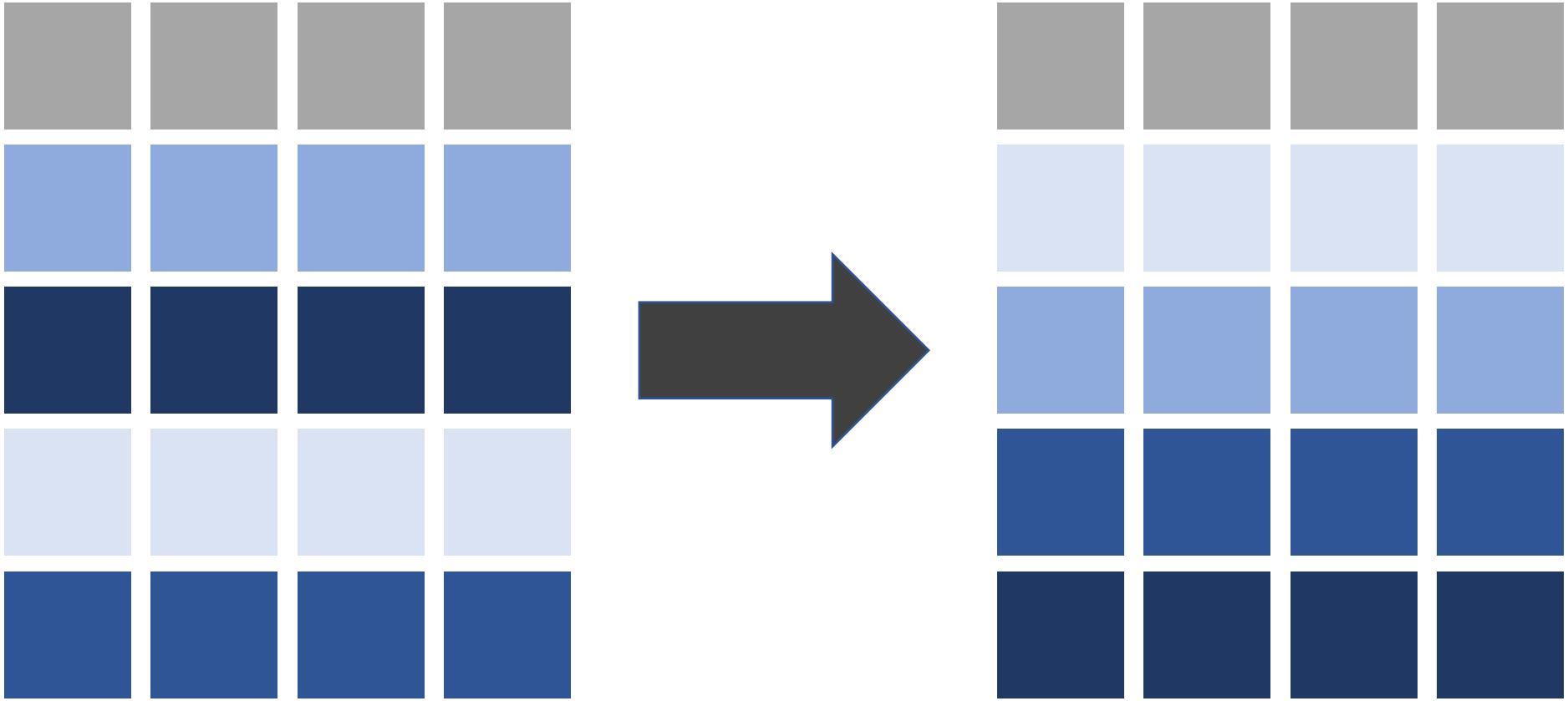
```
mutate(diamonds, price = price * 1.3)
mutate(diamonds, price_aud = price * 1.3)
mutate(diamonds, ppc = price/carat)
mutate(diamonds, colour = str_to_lower(colour))
```

filter()



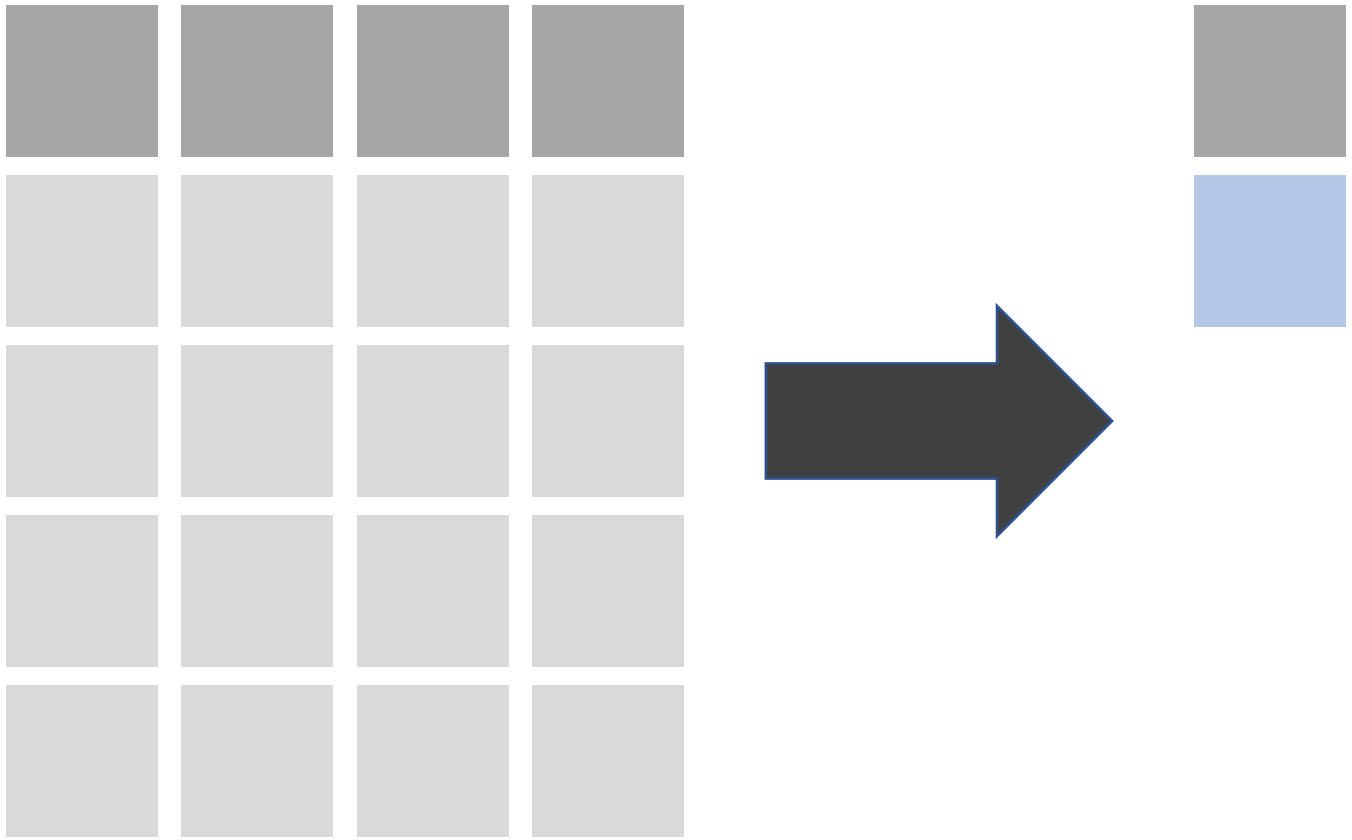
```
filter(diamonds, carat > 0.24, cut == ‘Good’)

filter(diamonds, colour %in% c(‘I’, ‘J’))
```

arrange()



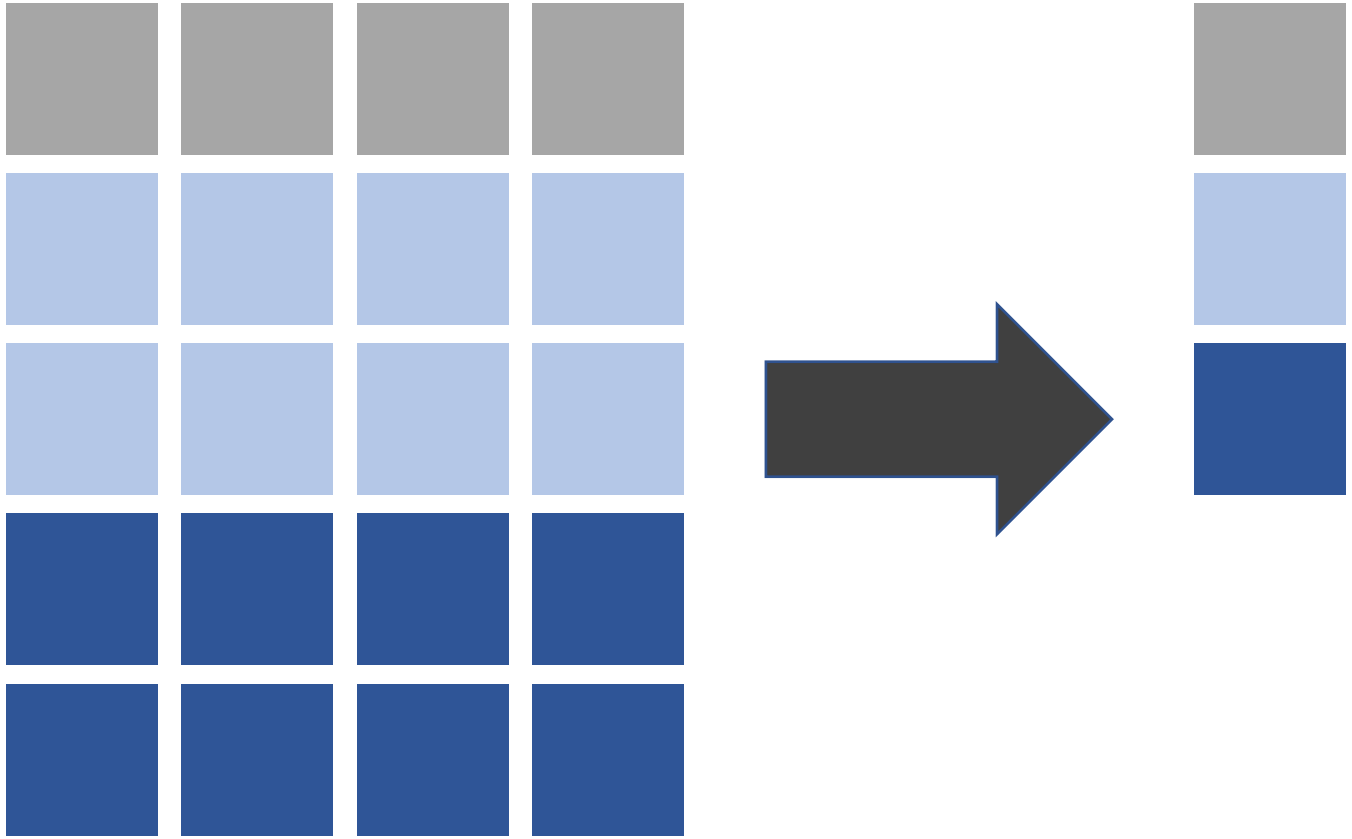
```
arrange(diamonds, cut, colour)

arrange(diamonds, desc(cut), colour)
```

summarise()



```
summarise(diamonds, m_carat = mean(carat))
```

group_by() then summarise()



```
summarise(group_by(diamonds, colour), m_price = mean(price))
```

```
by_colour <- group_by(diamonds, colour)
summarise(by_colour, m_price = mean(price))
```

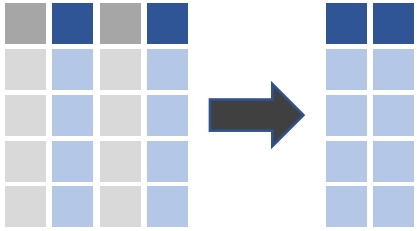
pipe

%>%

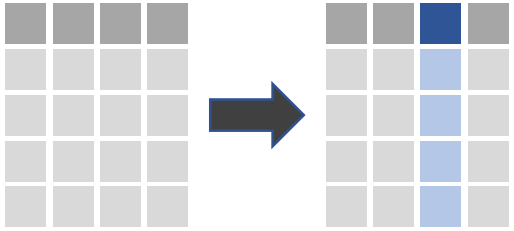
```
diamonds %>%
    group_by(colour) %>%
    summarise(m_price = mean(price))
```

dplyr

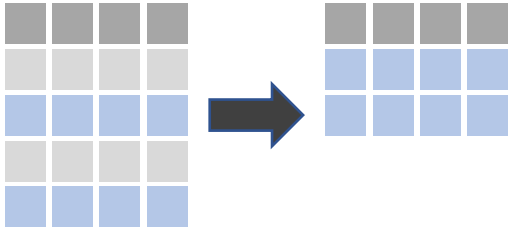
select()



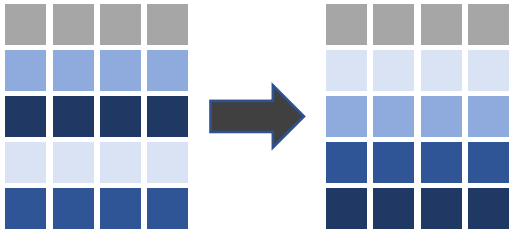
mutate()



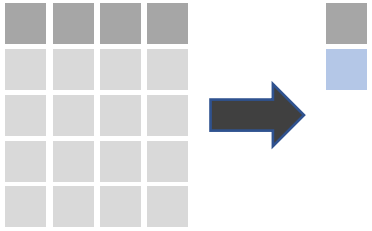
filter()



arrange()



summarise()



rapaport report

RAPAPORT : (.30 - .39 CT.) : 03/13/20											
	IF	VVS1	VVS2	VS1	VS2	SI1	SI2	SI3	I1	I2	I3
D	38	30	28	26	24	21	18	17	16	11	7
E	30	28	26	24	22	20	17	16	15	10	6
F	28	27	25	23	21	19	17	15	14	9	6
G	26	24	23	22	20	18	16	14	13	8	5
H	23	22	21	20	19	17	16	13	11	8	5
I	22	21	20	19	18	16	15	12	10	7	5
J	20	19	18	17	16	15	14	11	9	7	4
K	18	17	16	15	14	13	12	10	8	6	4
L	17	16	15	14	13	12	10	9	6	5	3
M	16	15	14	13	12	11	9	8	5	4	3

ROUNDS

RAPAPORT : (.40 - .49 CT.) : 03/13/20											
	IF	VVS1	VVS2	VS1	VS2	SI1	SI2	SI3	I1	I2	I3
D	47	37	34	31	29	25	21	20	18	12	8
E	38	35	32	29	27	23	20	19	17	11	7
F	35	33	30	27	26	22	19	18	16	11	7
G	31	29	27	26	25	21	19	17	15	10	6
H	27	26	25	24	23	20	18	16	14	9	6
I	25	24	23	22	21	19	18	15	13	8	6
J	22	21	20	19	18	17	16	14	12	8	5
K	20	19	18	17	16	15	14	12	10	7	5
L	18	17	16	15	14	13	12	10	8	6	4
M	17	16	15	14	13	12	11	9	7	5	4

- 0.45 carat
- F colour
- VVS1 clarity

price

$$= 0.45 \times \$3300$$

$$= \$1485$$

task 1 – prepare the data

a) Change the spelling of color column

b) We only want diamonds that are cut at least ‘Very Good’ and carat greater than equal to 0.3

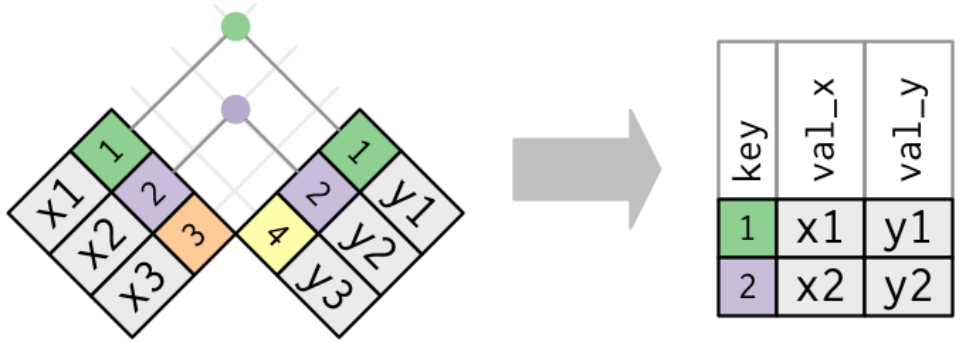
c) Add Price Per Carat (ppc) column in hundreds of $

d) Keep columns price, ppc, 4 C’s

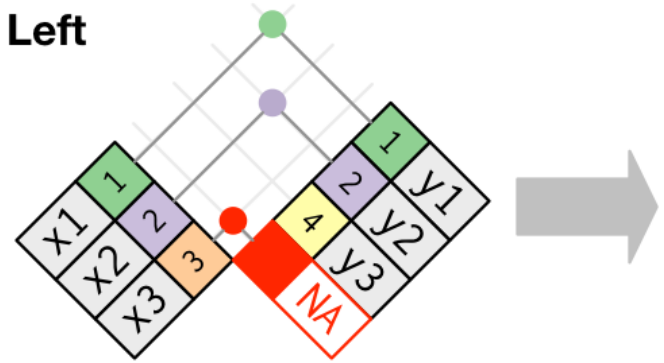
e) Assign this table to a new variable name

more dplyr

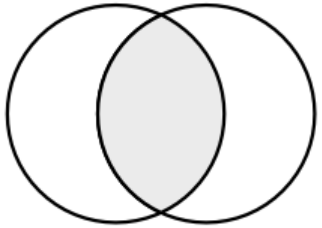
joins



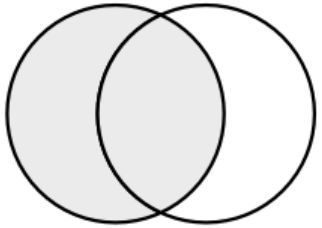
key	val_x	val_y
1	x1	y1
2	x2	y2



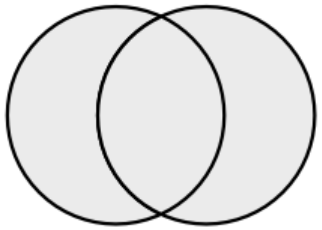
key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA



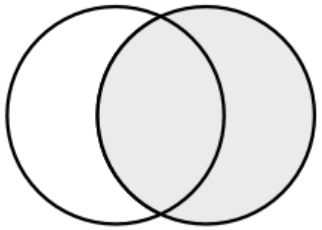
inner_join(x, y)



left_join(x, y)

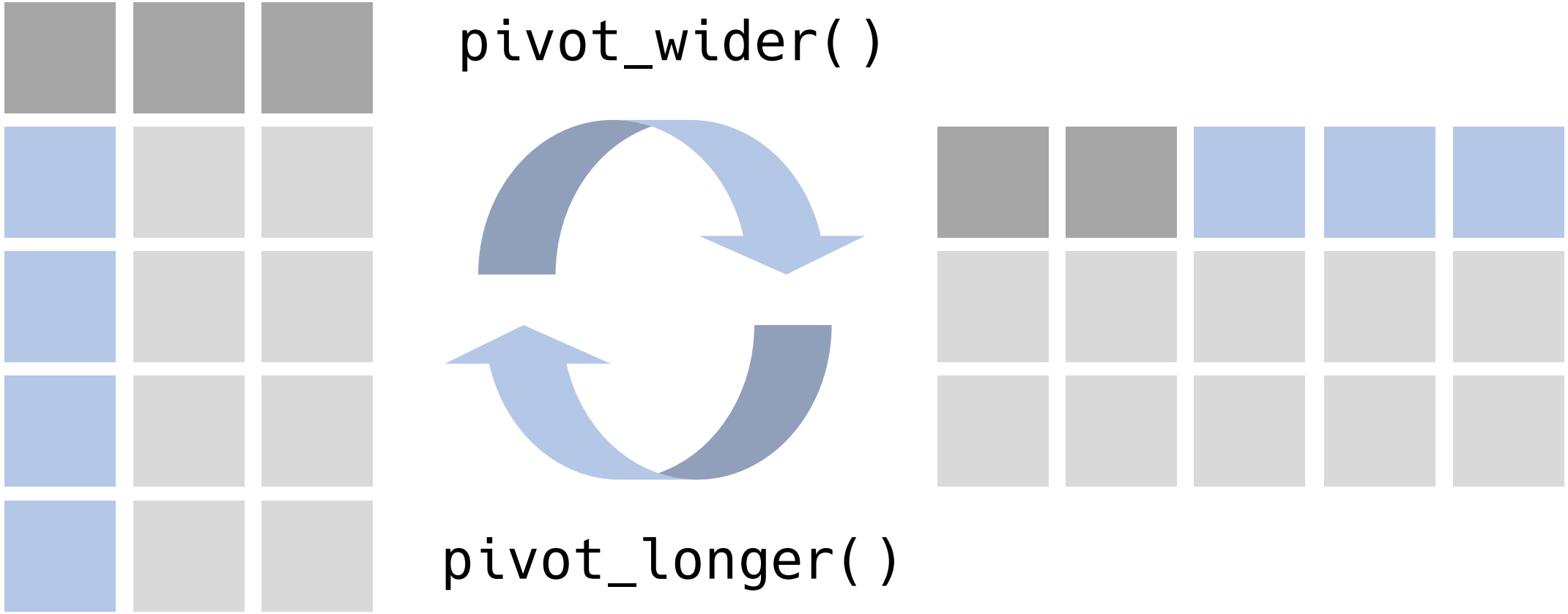


full_join(x, y)



right_join(x, y)

pivoting



rapaport report

RAPAPORT : (.30 - .39 CT.) : 03/13/20

	IF	VVS1	VVS2	VS1	VS2	SI1	SI2	SI3	I1	I2	I3
D	38	30	28	26	24	21	18	17	16	11	7
E	30	28	26	24	22	20	17	16	15	10	6
F	28	27	25	23	21	19	17	15	14	9	6
G	26	24	23	22	20	18	16	14	13	8	5
H	23	22	21	20	19	17	16	13	11	8	5
I	22	21	20	19	18	16	15	12	10	7	5
J	20	19	18	17	16	15	14	11	9	7	4
K	18	17	16	15	14	13	12	10	8	6	4
L	17	16	15	14	13	12	10	9	6	5	3
M	16	15	14	13	12	11	9	8	5	4	3

ROUNDS

RAPAPORT : (.40 - .49 CT.) : 03/13/20

	IF	VVS1	VVS2	VS1	VS2	SI1	SI2	SI3	I1	I2	I3
D	47	37	34	31	29	25	21	20	18	12	8
E	38	35	32	29	27	23	20	19	17	11	7
F	35	33	30	27	26	22	19	18	16	11	7
G	31	29	27	26	25	21	19	17	15	10	6
H	27	26	25	24	23	20	18	16	14	9	6
I	25	24	23	22	21	19	18	15	13	8	6
J	22	21	20	19	18	17	16	14	12	8	5
K	20	19	18	17	16	15	14	12	10	7	5
L	18	17	16	15	14	13	12	10	8	6	4
M	17	16	15	14	13	12	11	9	7	5	4

task 2 – create rapaport report

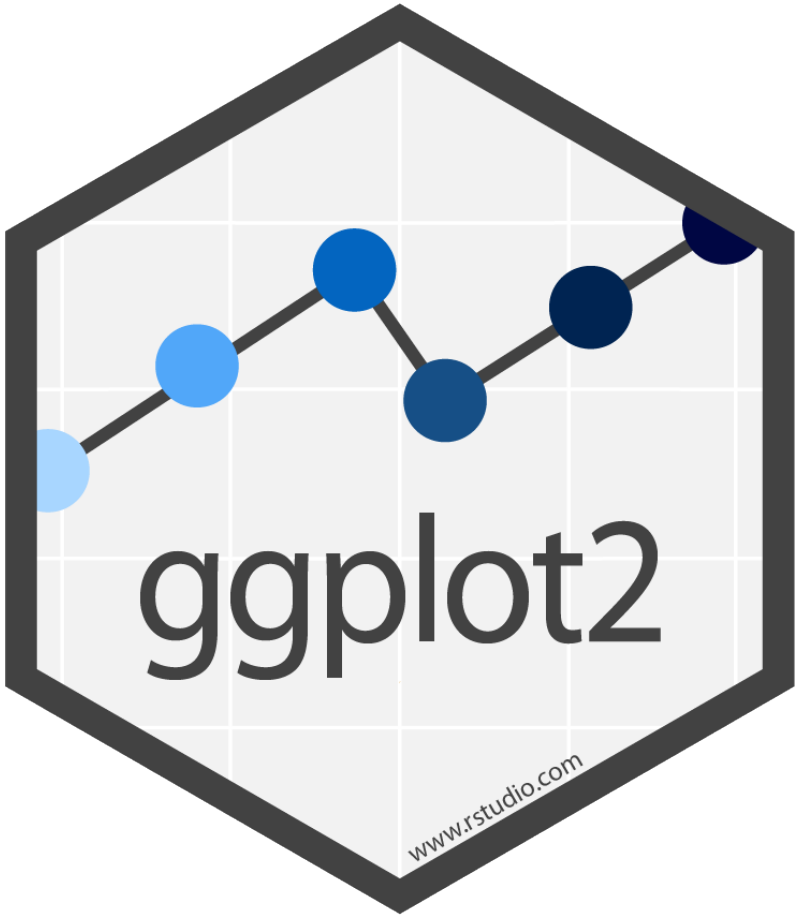
a) Add a column to result of task 1 using the cut function with the following intervals: (use ?cut to read the docs)

```
carat_lower_bounds <- c(0.3, 0.4, 0.5, 0.7, 0.9, 1, 1.5, 2, 3, 4, 5)
```

b) Find the median ppc within each carat interval, colour and clarity. Store in a new variable

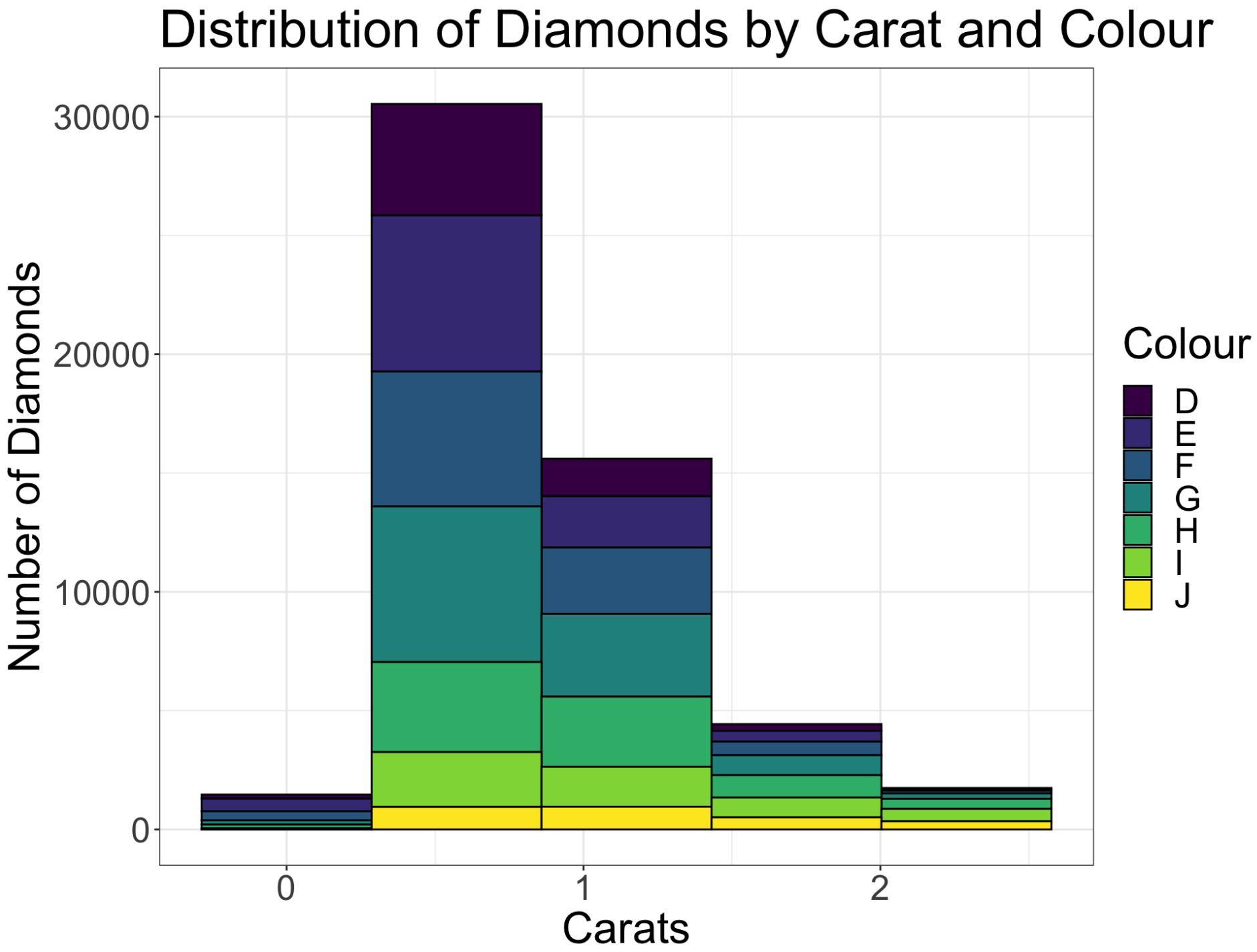
c) Create the Rapaport Report for diamonds with carat in [1, 1.5). Hint: Use pivot_wider()

ggplot2

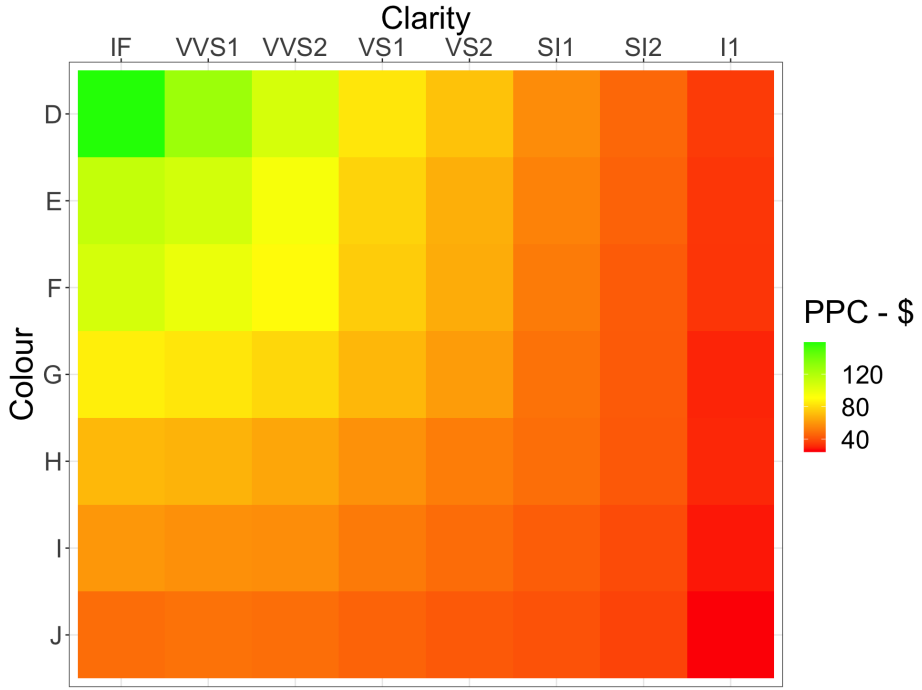


layers

```
ggplot(data, mapping = aes(x, y, fill, colour)) +

     geom_bar(), geom_histgram(), geom_point() +

     theme() +

     labs()
```


Distribution of Diamonds by Carat and Colour
Number of Diamonds
30000
20000
10000
0
0
1
2
Carats
Colour
D
E
F
G
H
I
J

task 3 – heat map



Hints
- Use geom_tile()
- Reorder the factors with reorder()

task 4 – analysis

a) What conclusions can we draw from our analysis?

b) Which diamond should I buy?

other awesome tools

a) shiny

b) rmarkdown

c) dbplyr (show query)

the best resource

R for Data Science, Hadley Wickham

https://r4ds.had.co.nz/

questions

thank you

ryan snoyman